Overview Dataset

"Data Pembuatan Semen (Cement Manufacturing)"

Beton merupakan material terpenting dalam teknik sipil. Kuat tekan beton adalah fungsi yang sangat nonlinier dari umur dan bahan. Bahan-bahan tersebut antara lain semen, terak tanur tinggi, fly ash, air, superplasticizer, agregat kasar, dan agregat halus. Dataset ini berisi data mengenai kekuatan semen, bahan penyusun dan waktu campuran.

Kekuatan tekan beton (MPa) untuk campuran tertentu di bawah umur tertentu (hari) ditentukan dari informasi laboratorium. Data ini merupakan data (tidak diskalakan). Data memiliki 8 variabel input kuantitatif, dan 1 variabel output kuantitatif, dan 1030 kejadian (pengamatan).

Pertemuan II

Pada praktikum ini, Anda akan melakukan beberapa operasi dasar statistik dengan data bertema keseshatan yaitu data tumor.

- Histogram
- Outliers
- Box Plot
- Summary Statistics
- Relationship Between Variables
- Correlation
- Covariance
- Pearson Correlation
- Spearman's Rank Correlation
- Mean VS Median
- Hypothesis Testing

Instruksi Praktikum

1. Silahkan modifikasi kode operasi yang ada menggunakan library perhitungan berbasis GPU (Library Cupy)
2. Bacalah dataset yang berada tersimpan url https://raw.githubusercontent.com/supasonicx/ATA-praktikum-01/main/concrete.csv
3. Periksa dataset apakah terdapat data yang bernilai null dengan menggunakan fungsi .isnull()
4. Buatlah sebuah histogram dari data kolom 'strength'.
5. Buatlah diagram boxplot dari dataset yang ada.
6. Hitung karakteristik statistik (standar deviasi, variance, mean, median) dari masing-masing kolom data.
7. Buatlah correlation map dari dataset tersebut.
8. Hitung covariance dari kolom data yang diminta
9. Hitung pearson correlation dan spearsman correlation dari kolom data yang diminta
10. Hitung nilai hipotesis testing untuk kolom age dan strength.

```
# import libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from pandas.plotting import autocorrelation_plot
from scipy import stats
plt.style.use("ggplot")
import warnings
warnings.filterwarnings("ignore")
from scipy import stats
```

Membaca Dataset

```
# read data as pandas data frame
url_data = "https://raw.githubusercontent.com/supasonicx/ATA-praktikum-01/main/concrete.csv"
data = pd.read_csv(url_data)
```

```
## Melihat 5 baris awal dari dataset yang digunakan
data.head()
```

	cement	slag	ash	water	superplastic	coarseagg	fineagg	age	strength
0	141.3	212.0	0.0	203.5	0.0	971.8	748.5	28	29.89
1	168.9	42.2	124.3	158.3	10.8	1080.8	796.2	14	23.51
2	250.0	0.0	95.7	187.4	5.5	956.9	861.2	28	29.22
3	266.0	114.0	0.0	228.0	0.0	932.0	670.0	28	45.85
4	154.8	183.4	0.0	193.3	9.1	1047.4	696.7	28	18.29

```
## Melihat dimensi dataset
print('Shape dataset', data.shape)
```

```
Shape dataset (1030, 9)
```

```
## Melihat kolom dataset
print(data.columns)
```

```
Index(['cement', 'slag', 'ash', 'water', 'superplastic', 'coarseagg',
       'fineagg', 'age', 'strength'],
      dtype='object')
```

```
print("mean stength :",data['strength'].isnull())
```

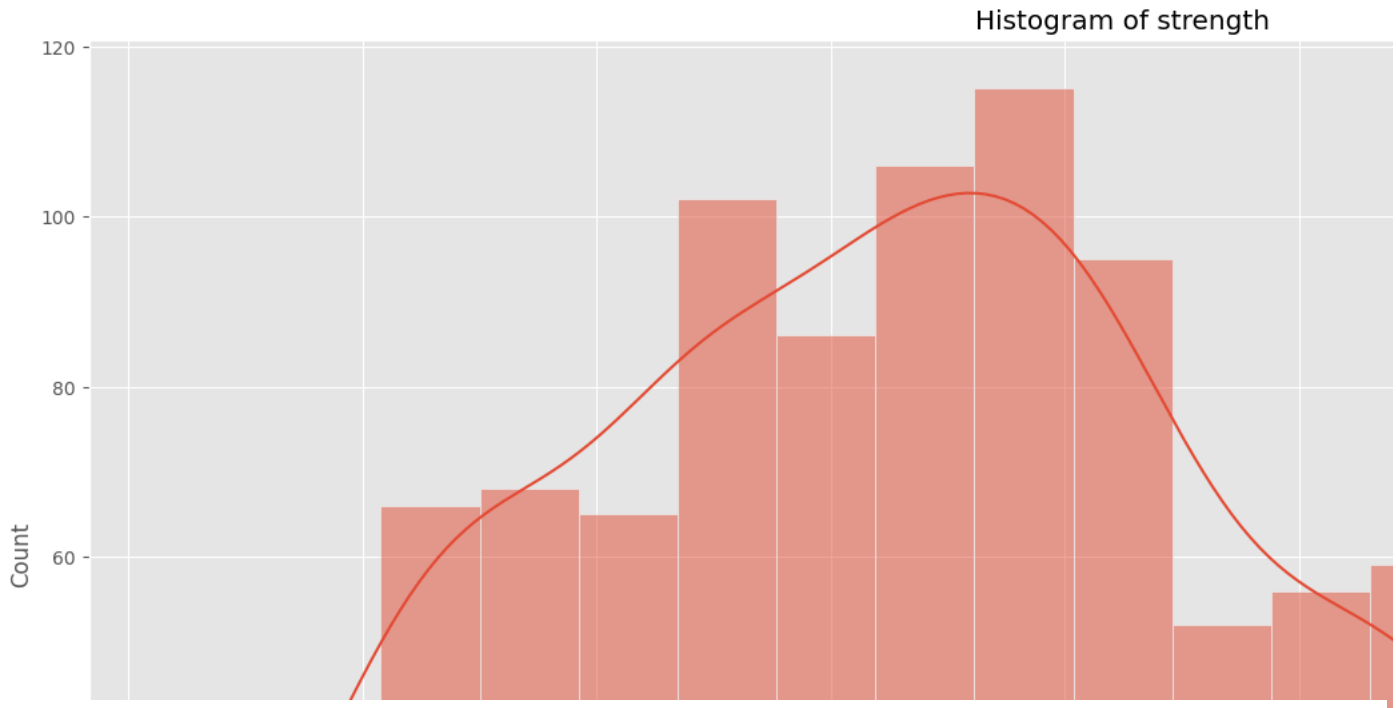
```
mean stength : 0       False
1       False
2       False
3       False
4       False
        ...
1025    False
1026    False
1027    False
1028    False
1029    False
Name: strength, Length: 1030, dtype: bool
```

Histogram

- Menampilkan Berapa kali (frekueni) setiap nilai muncul dalam kumpulan data.
- Jenis deskripsi ini disebut distribusi variabel
- Cara paling umum untuk merepresentasikan distribusi variabel adalah histogram yaitu grafik yang menunjukkan frekuensi dari setiap nilai.
- Frequency = berapa kali setiap nilai muncu
- Contoh: [1,1,1,1,2,2,2]. Frequency dari 1 adalah empat dan frequency dari 2 adalah tiga.

```
## Buatlah histogram dari kolom strength
plt.figure(figsize=(20,10))
plt.title('Histogram of strength')
sns.histplot(data,x='strength',kde=True)
plt.hist(data['strength'], density=True, bins=30, label="Data")
```

```
(array([0.00217712, 0.00399138, 0.01015989, 0.01269986, 0.02104548,
        0.01451412, 0.01088559, 0.01814265, 0.02757683, 0.01596553,
        0.01995692, 0.02975395, 0.01814265, 0.03047966, 0.02249689,
        0.01850551, 0.012337  , 0.00979703, 0.01451412, 0.01269986,
        0.00907133, 0.00798277, 0.00507994, 0.00580565, 0.00580565,
        0.00362853, 0.00362853, 0.00181427, 0.00362853, 0.00145141]),
 array([ 2.33      ,  5.00566667,  7.68133333, 10.357     , 13.03266667,
        15.70833333, 18.384     , 21.05966667, 23.73533333, 26.411     ,
        29.08666667, 31.76233333, 34.438     , 37.11366667, 39.78933333,
        42.465     , 45.14066667, 47.81633333, 50.492     , 53.16766667,
        55.84333333, 58.519     , 61.19466667, 63.87033333, 66.546     ,
        69.22166667, 71.89733333, 74.573     , 77.24866667, 79.92433333,
        82.6       ]),
 <BarContainer object of 30 artists>)
```



Box Plot

- Anda dapat melihat outlier juga dari box plot
- Temukan outlier pada dataset ini

```
## Buatlah histogram dari kolom strength
q1 = data[["strength"]].quantile(0.35)
q2 = data[["strength"]].quantile(0.85)
iqr = q2 -q1
print(iqr)
```

```
strength    26.099
dtype: float64
```

Summary Statistics

- Mean/rata-rata
- Variance: penyebaran distribusi
- Standart deviation square root dari variance
- Mari kita lihat ringkasan statistik rata-rata pancaran tumor jinak:

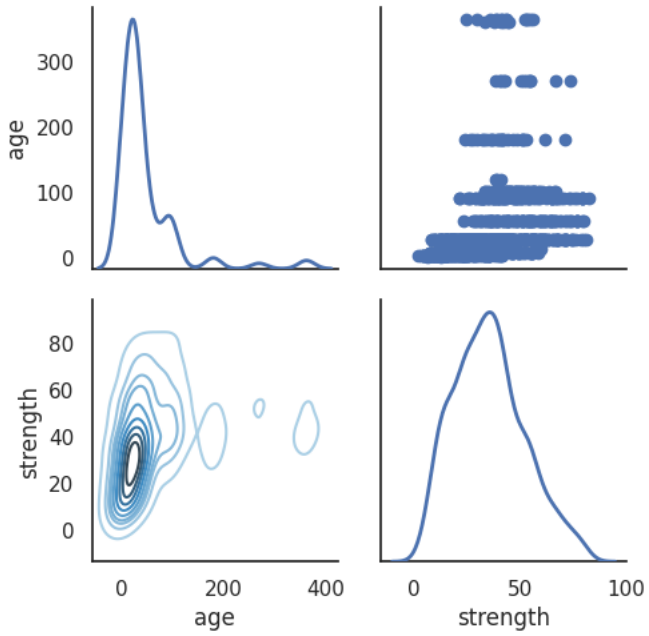
```
## Hitung karakteristik data dari masing-masing kolom dengan menggunakan perintah describe.
print("standart deviation :", data.strength.std())
print("variance           :", data.strength.var())
print("mean               :", data.strength.mean())
print("mdian              :", data.strength.median())
```

```
standart deviation : 16.705741961912512
variance           : 279.08181449800446
mean               : 35.817961165048544
mdian              : 34.445
```

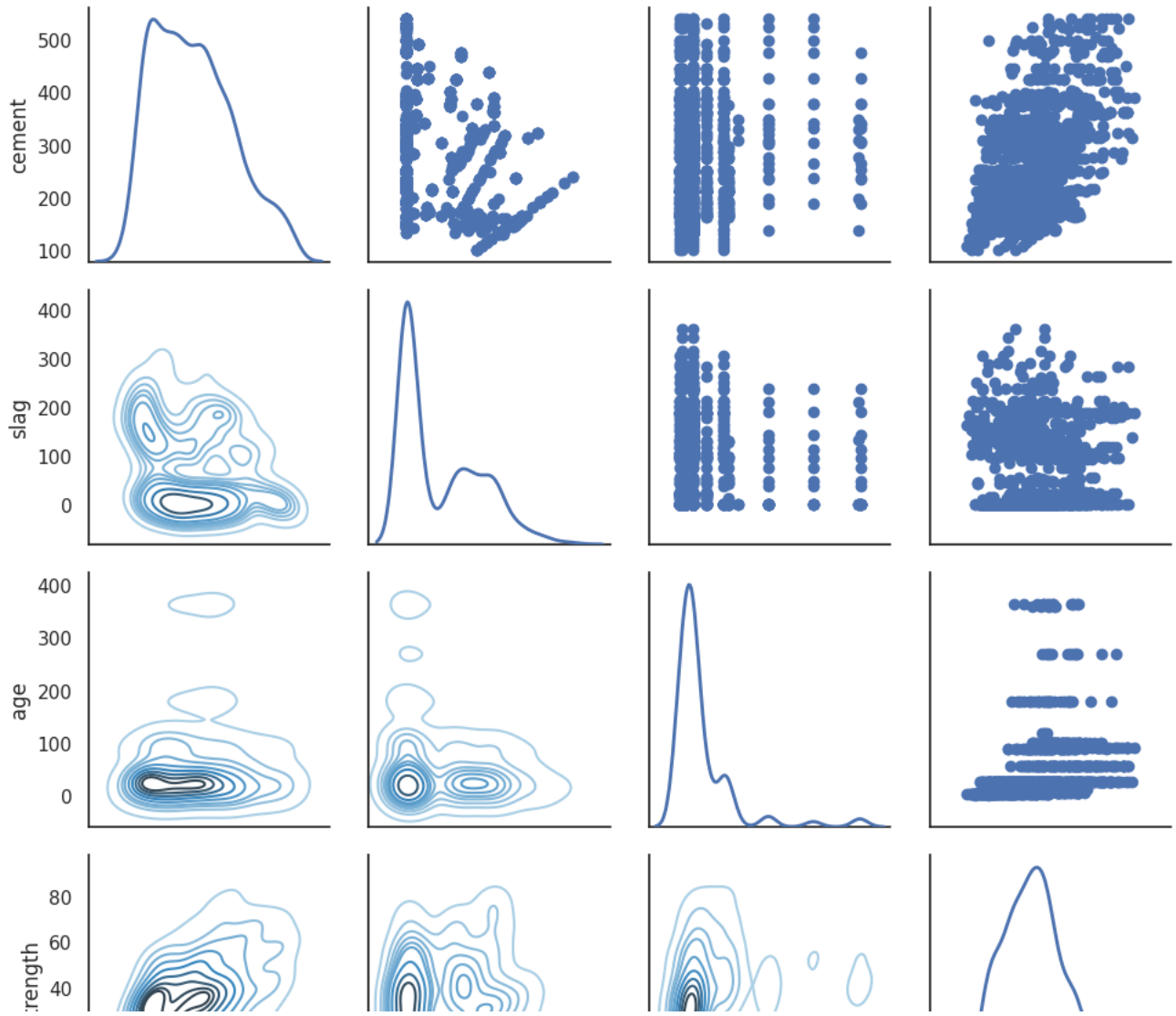
Relationship Between Variables

- Kita dapat mengatakan bahwa dua variabel terkait satu sama lain, jika salah satunya memberikan informasi tentang yang lain
- Misalnya, harga dan jarak. Jika Anda pergi jarak jauh dengan taksi Anda akan membayar lebih. Oleh karena itu kita dapat mengatakan bahwa harga dan jarak berhubungan positif satu sama lain.
- Scatter Plot, Cara termudah untuk memeriksa hubungan antara dua variabel
- Mari kita lihat hubungan antara radius mean dan mean area
- Di scatter plot Anda dapat melihat bahwa ketika radius mean meningkat, mean area juga meningkat. Oleh karena itu, mereka berkorelasi positif satu sama lain.
- Tidak ada korelasi antara mean area dan dimensi fraktal se. Karena ketika mean area berubah, dimensi fraktal se tidak terpengaruh oleh peluang mean area

```
# Tampilkan hubungan antara data kolom 'age' dan 'strength'
sns.set(style = "white")
df = data.loc[:,["age","strength"]]
g = sns.PairGrid(df,diag_sharey = False)
g.map_lower(sns.kdeplot,cmap="Blues_d")
g.map_upper(plt.scatter)
g.map_diag(sns.kdeplot,lw = 2)
plt.show()
```




```
# Tampilkan hubungan antara data kolom 'cement', 'slag','age' dan 'strength'
sns.set(style = "white")
df = data.loc[:,["cement","slag","age","strength"]]
g = sns.PairGrid(df,diag_sharey = False)
g.map_lower(sns.kdeplot,cmap="Blues_d")
g.map_upper(plt.scatter)
g.map_diag(sns.kdeplot,lw = 2)
plt.show()
```



Correlation

- Kekuatan hubungan antara dua variabel
- Mari kita lihat korelasi antara semua fitur.

```
## Buatlah diagram heatmap dari setiap kolom yang ada dengan library seaborn
f,ax=plt.subplots(figsize = (15,15))
sns.heatmap(data.corr(),annot= True,linewidths=0.5,fmt = ".1f" ,ax=ax)
plt.xticks(rotation=90)
plt.yticks(rotation=0)
plt.title('correlation Map')
plt.savefig('graph.png')
plt.show()
```

- Matriks besar yang mencakup banyak angka
- Kisaran angka ini adalah -1 hingga 1.
- Arti dari 1 adalah dua variabel yang saling berkorelasi positif seperti mean radius dan mean area
- Arti dari nol adalah tidak ada korelasi antara variabel seperti rata-rata radius dan fractal dimension se
- Arti dari -1 adalah dua variabel berkorelasi negatif satu sama lain seperti rata-rata radius dan mean/rata-rata fractal dimension.Sebenarnya korelasi antara keduanya bukan -1, melainkan -0,3 tetapi idenya adalah jika tanda korelasi negatif berarti ada adalah korelasi negatif.

Covariance

- Covariance adalah ukuran kecenderungan dua variabel untuk bervariasi bersama-sama
- Jadi covarians dimaksimalkan jika dua vektor identik
- Covarians adalah nol jika mereka ortogonal.
- Covariance negatif jika mereka menunjuk ke arah yang berlawanan

- Mari kita lihat kovarians antara mean radius dan mean area. Kemudian lihat radius mean dan fractal dimension se

```
## Bandingkan nilai covariance dari data age, strength dan strength dan cement
print("Covariance diantara radius mean dan area mean: ",data.strength.cov(data.age))
print("Covariance diantara radius mean dan fractal dimension se: ",data.strength.cov(data.cement))
```

```
Covariance diantara radius mean dan area mean:  347.05975751743136
Covariance diantara radius mean dan fractal dimension se:  869.1430218800419
```

Pearson Correlation

- Pembagian covarians dengan standar deviasi variabel
- Mari kita lihat korelasi pearson antara mean/rata-rata radius dan mean/rata-rata area
- Pertama mari kita gunakan metode .corr() yang sebenarnya kita gunakan pada bagian korelasi. Di bagian korelasi kami sebenarnya menggunakan korelasi pearson :)
- p1 dan p2 adalah sama. Di p1 kita menggunakan metode corr(), di p2 kita menerapkan definisi korelasi pearson (cov(A,B)/(std(A)*std(B)))
- Seperti yang kita harapkan korelasi pearson antara area_mean dan area_mean adalah 1 yang berarti bahwa mereka adalah distribusi yang sama
- Juga pearson correlation antara area_mean dan radius_mean adalah 0,98 yang berarti saling berkorelasi positif dan hubungan antar keduanya sangat tinggi.
- Untuk lebih jelas apa yang kami lakukan di bagian korelasi dan bagian korelasi pearson adalah sama.

```
## Hitung nilai pearson correlation dari kolom data cement dan age
p1 = data.loc[:,["age","cement"]].corr(method= "pearson")
p2 = data.cement.cov(data.age)/(data.cement.std()*data.age.std())
print('Pearson correlation: ')
print(p1)
print('Pearson correlation: ' ,p2)
```

```
Pearson correlation:
            age    cement
age     1.000000  0.081946
cement  0.081946  1.000000
Pearson correlation:  0.08194602387182238
```

Spearman's Rank Correlation

- Pearson correlation bekerja dengan baik jika hubungan antara variabel linier dan variabel kira-kira normal. Tapi itu tidak kuat, jika ada outlier
- Untuk menghitung korelasi spearman, kita perlu menghitung peringkat dari setiap nilai

```
## Hitung nilai spearsman rank dari kolom data age dan strength
ranked_data = data.rank()
spearman_corr = ranked_data.loc[:,["age","strength"]].corr(method= "pearson")
print("Spearman's correlation: ")
print(spearman_corr)
```

```
Spearman's correlation:
               age  strength
age       1.000000  0.596028
strength  0.596028  1.000000
```

- Korelasi Spearman sedikit lebih tinggi dari korelasi pearson
  - Jika hubungan antar distribusi tidak linier, korelasi spearman cenderung lebih baik dalam memperkirakan kekuatan hubungan
  - Korelasi Pearson dapat dipengaruhi oleh outlier, sehingga jika data Anda memiliki outlier, maka teknik Korelasi Spearman's Rank dapat digunakan.

Hypothesis Testing

- Classical Hypothesis Testing / Pengujian Hipotesis Klasik
- Apa yang Anda perlu lakukan untuk menjawab pertanyaan berikut : "diberikan sampel dan efek nyata, berapa peluang melihat efek seperti itu secara kebetulan"
- Langkah pertama adalah mengukur ukuran efek nyata dengan memilih statistik uji. Pilihan alami untuk statistik uji adalah perbedaan mean/rata-rata antara dua kelompok.
- Langkah kedua adalah mendefinisikan hipotesis nol yaitu model sistem berdasarkan asumsi bahwa efek yang tampak tidak nyata. Hipotesis nol adalah jenis hipotesis yang digunakan dalam statistik yang menyatakan bahwa tidak ada signifikansi statistik dalam serangkaian pengamatan yang diberikan. Hipotesis nol adalah hipotesis yang orang mencoba untuk menyangkalnya. Hipotesis alternatif adalah hipotesis yang orang ingin mencoba untuk membuktikannya.
- Langkah ketiga adalah menghitung p-value yaitu probabilitas melihat efek nyata jika hipotesis nol benar. Misalkan kita memiliki uji hipotesis nol. Kemudian kita hitung nilai p. Jika nilai p kurang dari atau sama dengan ambang batas, kami menolak hipotesis nol.
- Jika p-value rendah, pengaruh tersebut dikatakan signifikan secara statistik artinya tidak mungkin terjadi secara kebetulan. Oleh karena itu kita dapat mengatakan bahwa efeknya lebih mungkin muncul pada populasi yang lebih besar.
- Mari kita coba contohkan. Hipotesis nol: dunia rata. Hipotesis alternatif: dunia itu bulat. Beberapa ilmuwan mulai menyangkal hipotesis nol. Ini akhirnya mengarah pada refleksi hipotesis nol dan penerimaan hipotesis alternatif.
- Contoh lainnya. "efek ini nyata" ini adalah hipotesis nol. Berdasarkan asumsi itu kami menghitung probabilitas efek yang tampak. Itu adalah nilai-p. Jika nilai p rendah, kami menyimpulkan bahwa hipotesis nol tidak mungkin benar.
- Sekarang mari kita buat contoh kita:
    - Saya ingin mengetahui apakah rata-rata radius dan rata-rata area terkait satu sama lain? Hipotesis nol saya adalah bahwa "hubungan antara rata-rata radius dan rata-rata area adalah nol pada populasi tumor'.
    - Sekarang kita perlu menyangkal hipotesis nol ini untuk menunjukkan bahwa mean/rata-rata radius dan mean/rata-rata area berhubungan. (walaupun sebenarnya kita telah mengetahui hasilnya berdasarkan analisa korelasi yang telah dilakukan sebelumnya)
    - mari kita cari nilai p (nilai probabilitas)

```
## Lakukan hubungan hipotesis data antara kolom age dan strength
statistic, p_value = stats.ttest_rel(data.age,data.strength)
print('HIPOTESIS AGE DAN STRENGHT adalah: ',p_value)
```

```
HIPOTESIS AGE DAN STRENGHT adalah:  1.545311719208927e-07
```

- Hasil perhitungan P values/ Nilai P hampir mendekati nol, sehingga kita dapat menolak hipotesis nol. Penolakan hipotesis ini memiliki arti nilai rata-rata age dan rata-rata strength pada data ini saling berpengaruh.