MK Praktikum Unggulan Universitas Gunadarma

Nama Mata Kuliah: Praktikum Komputasi Big Data (Tingkat 2)

NAMA : Agrieva Xananda Pramuditha

Kelas : 2IA18

NPM : 50423070

Self Study

Sesuai namanya, moda self-study dimana anda akan menjalankan program pada perangkat masing-masing. Materi akan diberikan melalui platform virtual class.

Pertemuan II

Pada praktikum ini, Anda akan melakukan beberapa operasi dasar statistik dengan data bertema konstruksi yaitu data pembuatan semen (Cement Manufacturing).

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

Materi Praktikum Dikhususkan Untuk Program Studi:

Fakultas Teknologi Industri (FTI)

No	Program Studi	Jenjang
1	Informatika	Sarjana

No	Program Studi	Jenjang
2	Teknik Elektro	Sarjana

Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI)

No	Program Studi	Jenjang
1	Sistem Informasi	Sarjana
2	Sistem Komputer	Sarjana
3	Manajemen Informatika	Diploma Tiga
4	Teknik Komputer	Diploma Tiga

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).

Dataset ini berisikan beberapa kategori sebagai berikut :

- cement (kg)
- slag (blast furnace slag, kg)
- ash (fly ash, kg)
- water (kg)
- superplastic (superplasticizer, kg)
- coarseagg (coarse aggregate, kg)
- fineagg (fine aggregate, kg)
- age (days, 1-365)
- strength (Concrete compressive strength, MPa)

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'`.

JAWABAN

Catatan!
Pada bagian #CODE HERE, <...> digantikan dengan code python sesuai dengan perintah.

Klik dua kali (atau tekan Enter) untuk mengedit

Import Libraries

```
# 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

```
# Membaca data sebagai pandas data frame
# CODE HERE
url = "https://raw.githubusercontent.com/supasonicx/ATA-praktikum-01/main/concret
data = pd.read_csv(url)


# 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
data.shape
```

```
(1030, 9)
```

```
# Melihat kolom dataset
data.columns
```

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

```
# mencetak nilai rata rata kekuatan semen pada dataset
print("mean strength :",data['strength'].mean())
```

```
mean strength : 35.817961165048544
```

```
# Periksa dataset apakah terdapat data yang bernilai null dengan menggunakan fung
# CODE HERE
print("Apakah dataset terdapat nilai Null : \n", data.isnull())
```

```
Apakah data Null :
       cement   slag    ash  water  superplastic  coarseagg  fineagg    age  \
0      False  False  False  False         False      False    False  False
1      False  False  False  False         False      False    False  False
2      False  False  False  False         False      False    False  False
3      False  False  False  False         False      False    False  False
4      False  False  False  False         False      False    False  False
...      ...    ...    ...    ...           ...        ...      ...    ...
1025   False  False  False  False         False      False    False  False
1026   False  False  False  False         False      False    False  False
1027   False  False  False  False         False      False    False  False
1028   False  False  False  False         False      False    False  False
1029   False  False  False  False         False      False    False  False

      strength
0        False
1        False
2        False
3        False
4        False
...        ...
1025     False
```

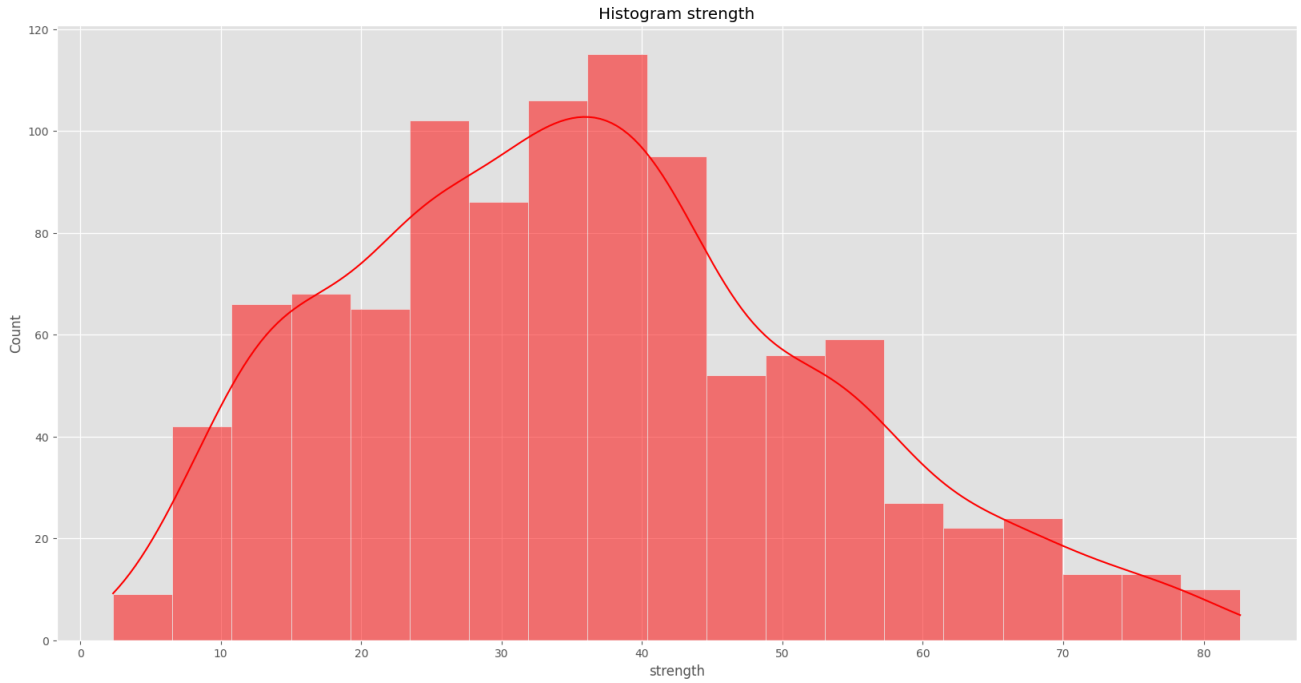
```
1026     False
1027     False
1028     False
1029     False

[1030 rows x 9 columns]
```

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'
# CODE HERE
plt.figure(figsize=(20,10))
plt.title("Histogram strength")
sns.histplot(data,x= "strength", kde = True, color='red', fill=True)
```

```
<Axes: title={'center': 'Histogram strength'}, xlabel='strength',
ylabel='Count'>
```



Box Plot

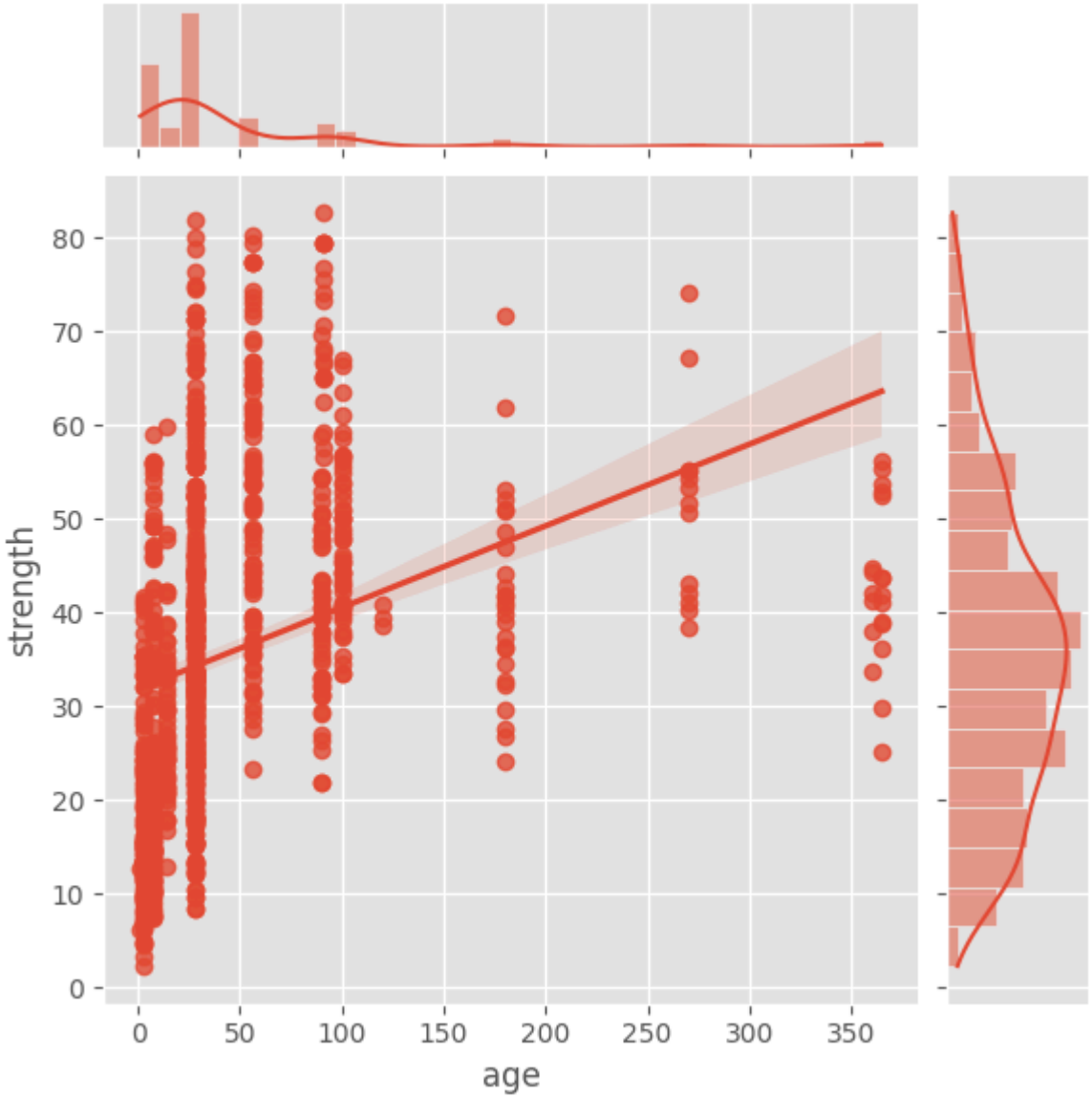
- Anda dapat melihat outlier juga dari box plot
- Temukan outlier pada dataset ini

```
# Buatlah Box Plot dari kolom 'strength'
# CODE HERE
Q1 = data['strength'].quantile(0.35)
Q3 = data['strength'].quantile(0.85)
IQR = Q3 - Q1

print(f'IQR (Interquartile Range) dari strength adalah: {IQR:.2f}')
```

```
IQR (Interquartile Range) dari strength adalah: 26.10
```

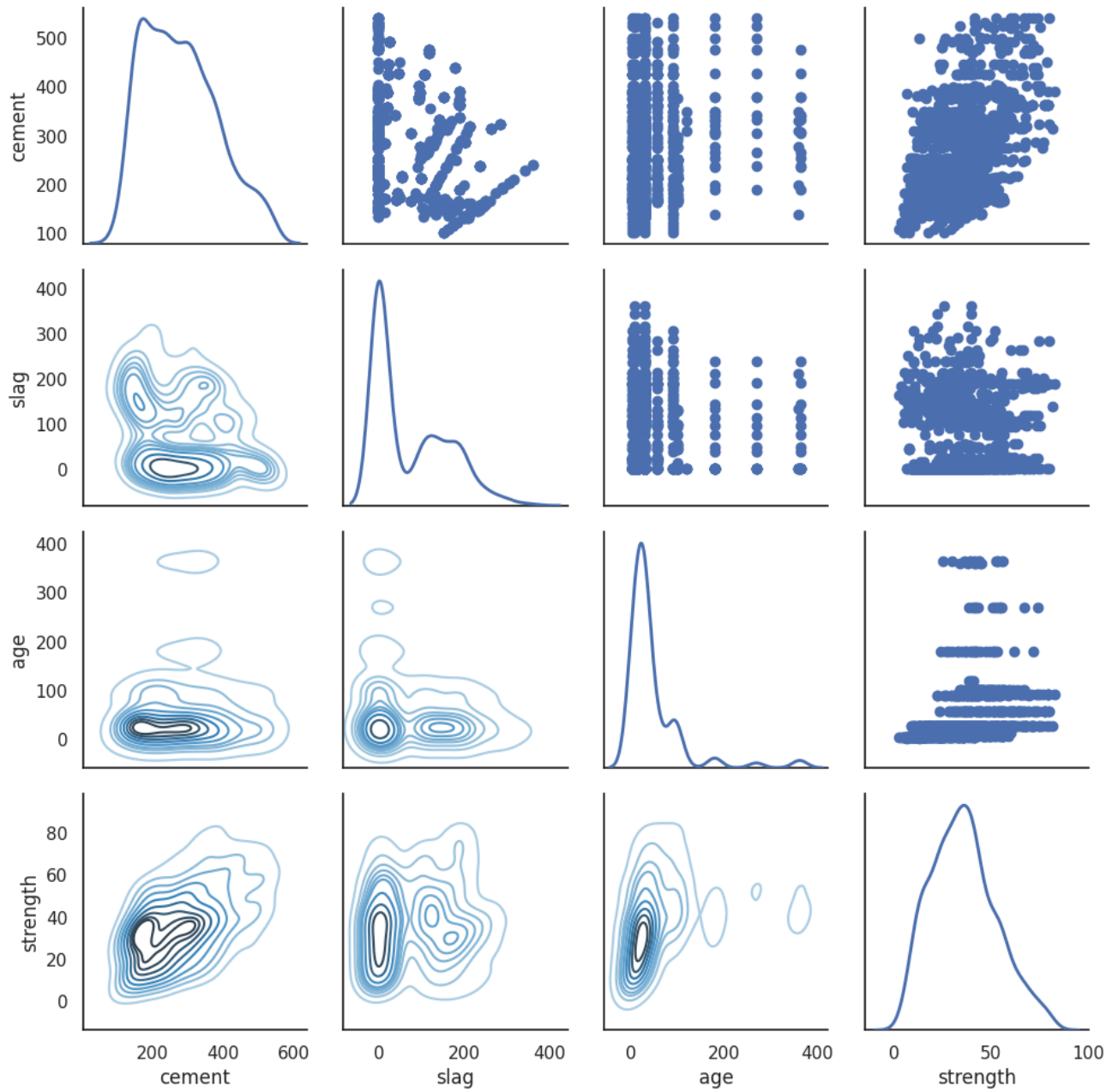
```
plt.figure(figsize = (20,10))
sns.jointplot(data = data, x = "age", y= "strength" ,kind="reg")
plt.show()
```

```
<Figure size 2000x1000 with 0 Axes>
```




```
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()
```



Summary Statistics

- Mean/rata-rata
- Variance: penyebaran distribusi
- Standart deviation square root dari variance
- Mari kita lihat ringkasan statistik pada data pembuatan semen (Cement Manufacturing):

```
# Hitung karakteristik data dari masing-masing kolom dengan menggunakan perintah
# CODE HERE

print("mean :", data.strength.mean())
print("variance :",data.strength.var())
print("standart deviation :", data.strength.std())
print("median :", data.strength.median())
```

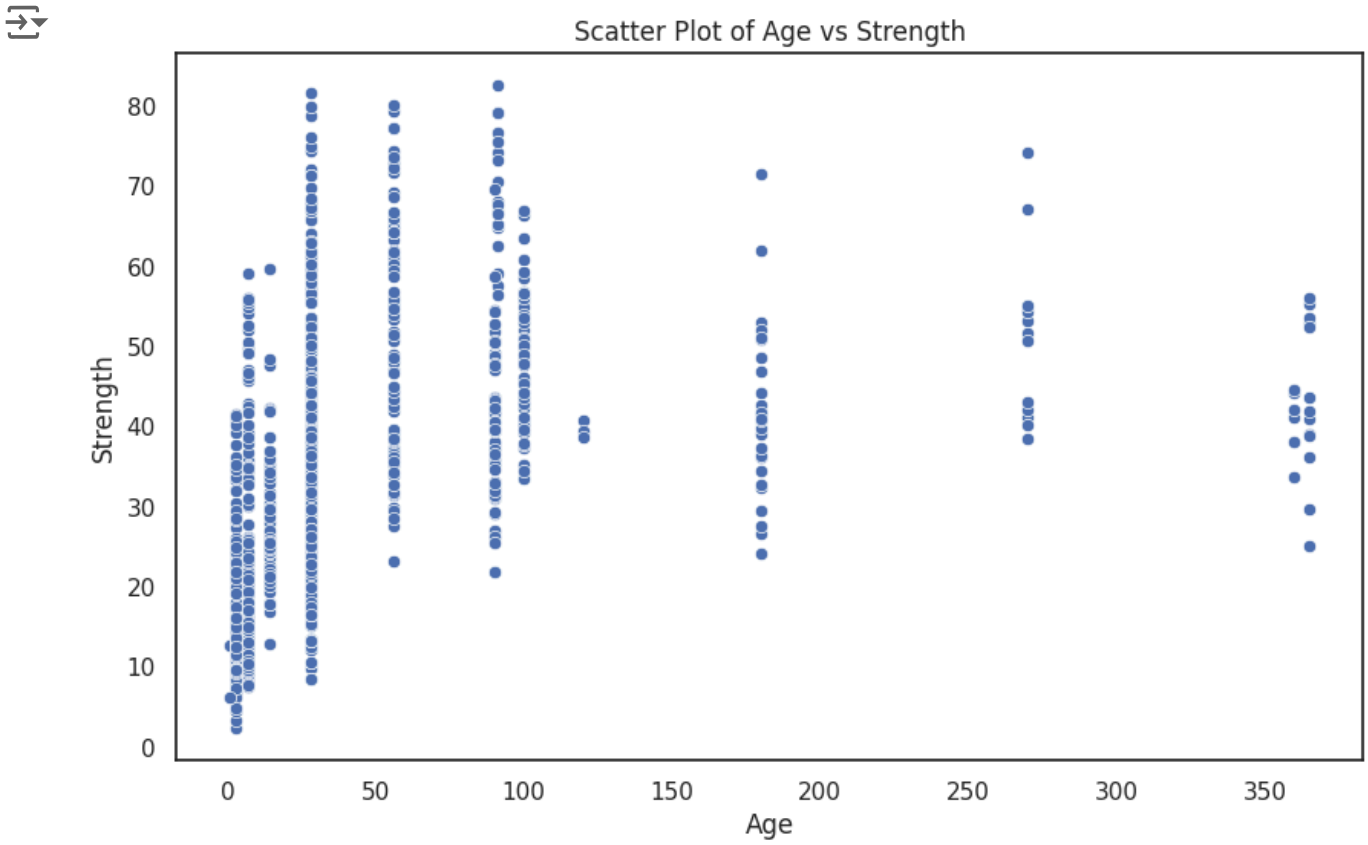
```
mean : 35.817961165048544
variance : 279.08181449800435
standart deviation : 16.70574196191251
median : 34.445
```

Relationship Between Variables

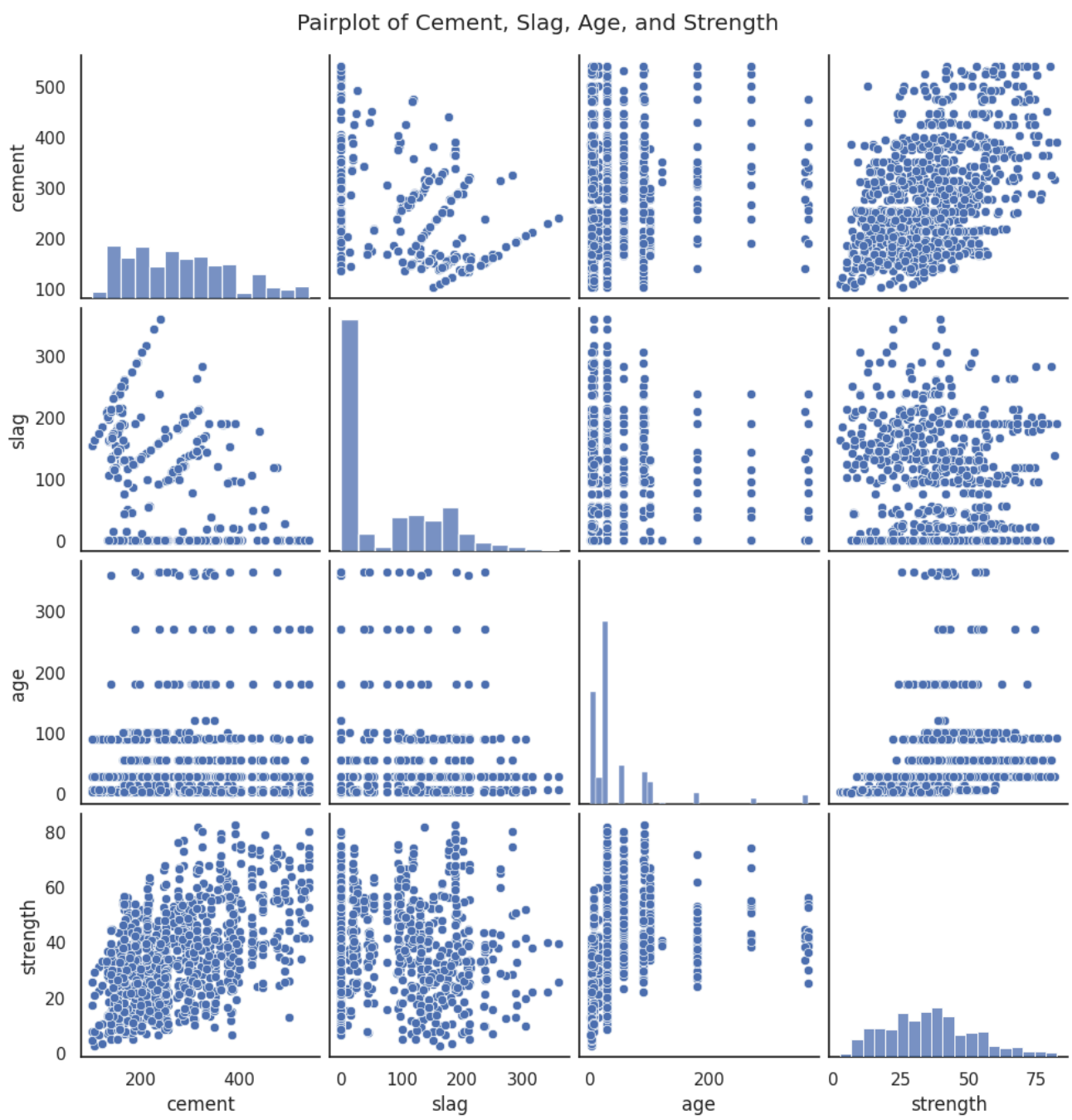
- Kita dapat mengatakan bahwa dua variabel terkait satu sama lain, jika salah satunya memberikan informasi tentang yang lain
- Misalnya, harga dan jarak. Jika Anda pergi jarak jauh dengan taksi Anda akan membayar lebih. Oleh karena itu kita dapat mengatakan bahwa harga dan jarak berhubungan positif satu 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'
# CODE HERE

plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='age', y='strength')
plt.title('Scatter Plot of Age vs Strength')
```

```
plt.xlabel('Age')
plt.ylabel('Strength')
plt.show()
```



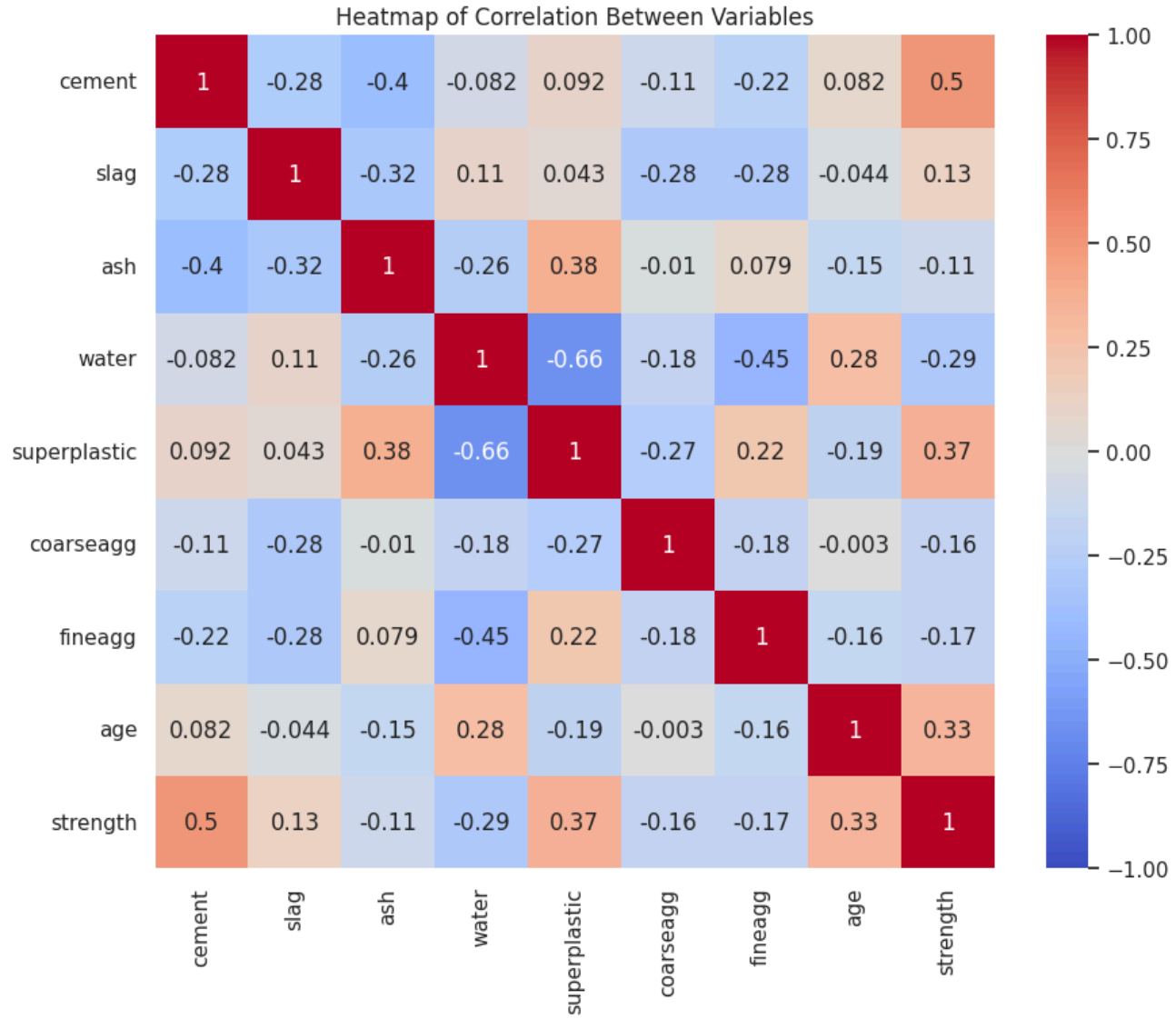

```
# Tampilkan hubungan antara data kolom 'cement', 'slag','age' dan 'strength'
# CODE HERE

sns.pairplot(data[['cement', 'slag', 'age', 'strength']])
plt.suptitle('Pairplot of Cement, Slag, Age, and Strength', y=1.02)
plt.show()
```


Pairplot of Cement, Slag, Age, and Strength
cement
slag
age
strength
0
100
200
300
400
500
0
100
200
300
0
100
200
300
0
20
40
60
80
200
400
0
100
200
300
0
200
0
25
50
75

Correlation

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

```
# Buatlah diagram heatmap dari setiap kolom yang ada dengan library seaborn
# CODE HERE

plt.figure(figsize=(10, 8))
correlation_matrix = data.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Heatmap of Correlation Between Variables')
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 diantara kolom 'strength' dengan kolom 'age' dan
# CODE HERE

cov_strength_age = np.cov(data['strength'], data['age'])[0, 1]
cov_strength_cement = np.cov(data['strength'], data['cement'])[0, 1]

cov_strength_age, cov_strength_cement
```

```
(347.05975751743136, 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 'cement' dan 'age
# CODE HERE
```

```
pearson_corr_cement_age = data['cement'].corr(data['age'])
cov_cement_age = np.cov(data['cement'], data['age'])[0, 1]
std_cement = data['cement'].std()
std_age = data['age'].std()
pearson_corr_manual = cov_cement_age / (std_cement * std_age)
```

```
(0.08194602387182238, 0.08194602387182195)
```

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'
# CODE HERE

spearman_corr_age_strength = data['age'].corr(data['strength'], method='spearman'
spearman_corr_age_strength
```

```
0.5960276340337732
```

- 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