Technical Test CGT Program Batch 2

Data Engineer

naufal.nashif

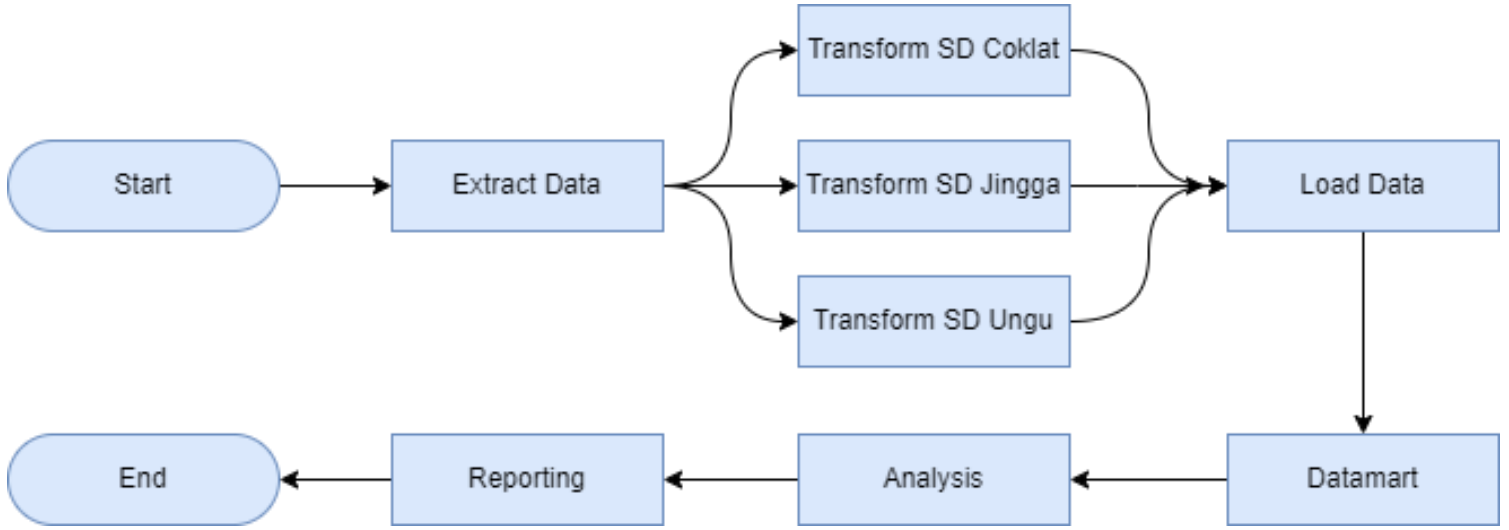
DATA ENGINEER

Naufal Nashif Imanuddin

Statistics and Data Science
IPB University

METHODOLOGY

Raw-data : SD Coklat.xlsx, SD Jingga.xlsx, SD Ungu.xlsx
Tools : Python - Collab Notebook, Excel

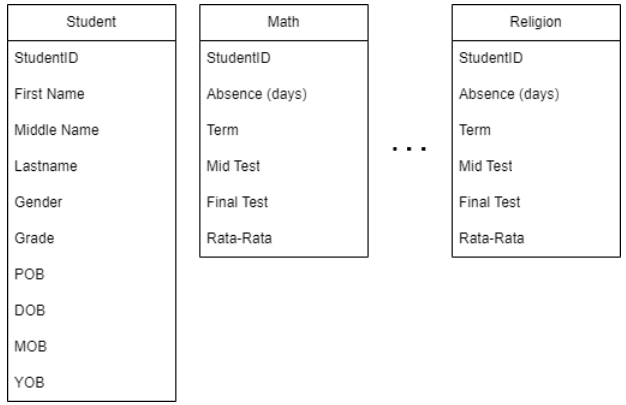


flowchart

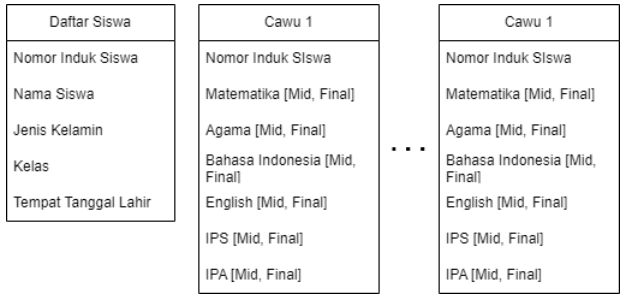
EXTRACT DATA

Raw-data : SD Coklat.xlsx, SD Jingga.xlsx, SD Ungu.xlsx
Tools : Python – Collab Notebook, Excel

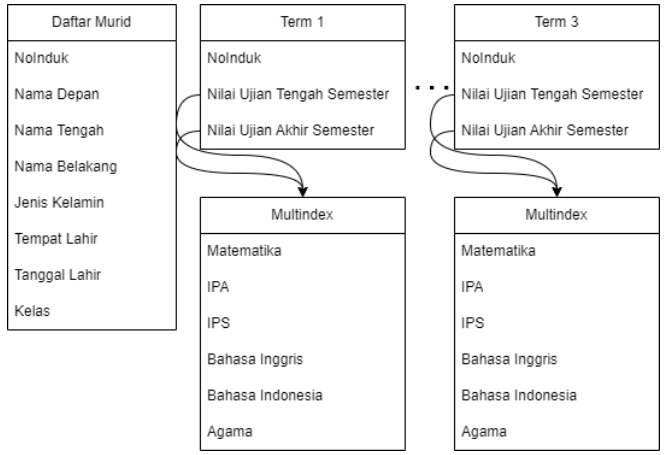
SD Coklat



SD Jingga



SD Ungu



TRANSFORM DATA

Dataset : SD Coklat.xlsx
Tools : Python – Collab Notebook

NIS	school	fullName	gender	grade	POB	birth	age	caturWulan	subjects	midTest	finalTest	mean	finalScore
Std001	SD Coklat	Ika Tiara Alatas	F	6	Pariaman	2011-01-05	13	1	Math	95	82	88.5	A
Std001	SD Coklat	Ika Tiara Alatas	F	6	Pariaman	2011-01-05	13	2	Math	66	76	71.0	B
Std001	SD Coklat	Ika Tiara Alatas	F	6	Pariaman	2011-01-05	13	3	Math	59	84	71.5	B

Kolom baru :school, fullName, birth, age, subjects, mean, finalScore

Kolom updated :-

Kolom deleted :First Name, Middle Name, Lastname, DOB, MOB, YOB, Absence (days)

TRANSFORM DATA

Dataset : SD Jingga.xlsx
Tools : Python – Collab Notebook, Excel

Excel : Merubah Multiindex menjadi tabular seperti dataset SD Coklat

Python :

NIS	school	fullName	gender	grade	POB	birth	age	caturWulan	subjects	midTest	finalTest	mean	finalScore
Std001	SD Jingga	Julie Helena Saad	F	6	Subulussalam	2011-04-02	12	1	Math	91	74	82.5	B
Std001	SD Jingga	Julie Helena Saad	F	6	Subulussalam	2011-04-02	12	2	Math	98	68	83.0	B
Std001	SD Jingga	Julie Helena Saad	F	6	Subulussalam	2011-04-02	12	3	Math	96	68	82.0	B

NIS	school	fullName	gender	grade	POB	birth	age	caturWulan	subjects	midTest	finalTest	mean	finalScore
Std001	SD Jingga	Julie Helena Saad	F	6	Subulussalam	2011-04-02	12	1	Math	91	74	82.5	B
Std001	SD Jingga	Julie Helena Saad	F	6	Subulussalam	2011-04-02	12	2	Math	98	68	83.0	B
Std001	SD Jingga	Julie Helena Saad	F	6	Subulussalam	2011-04-02	12	3	Math	96	68	82.0	B

Kolom baru :school, POB,birth, age, subjects, mean, finalScore

Kolom updated :gender (laki laki/perempuan), grade (romawi)

Kolom deleted :Tempat Tanggal Lahir

TRANSFORM DATA

Dataset : SD Ungu.xlsx
Tools : Python – Collab Notebook, Excel

Excel : Merubah Multiindex menjadi tabular seperti dataset SD Coklat

Python :

NIS	school	fullName	gender	grade	POB	birth	age	caturWulan	subjects	midTest	finalTest	mean	finalScore
Std001	SD Ungu	Mirza Patria Budiman	M	6	Tangerang	2009-09-18	14	1	Math	90	67	78.5	B
Std001	SD Ungu	Mirza Patria Budiman	M	6	Tangerang	2009-09-18	14	2	Math	51	97	74.0	B
Std001	SD Ungu	Mirza Patria Budiman	M	6	Tangerang	2009-09-18	14	3	Math	99	61	80.0	B

NIS	school	fullName	gender	grade	POB	birth	age	caturWulan	subjects	midTest	finalTest	mean	finalScore
Std001	SD Ungu	Mirza Patria Budiman	M	6	Tangerang	2009-09-18	14	1	Math	90	67	78.5	B
Std001	SD Ungu	Mirza Patria Budiman	M	6	Tangerang	2009-09-18	14	2	Math	51	97	74.0	B
Std001	SD Ungu	Mirza Patria Budiman	M	6	Tangerang	2009-09-18	14	3	Math	99	61	80.0	B

```
Kolom baru    :school, fullName, age, subjects, mean, finalScore
Kolom updated :gender (P/W), birth (12 nopember 2009)
Kolom deleted :Nama Depan, Nama Tengah, Nama Belakang
```

LOAD DATA

Dataset : Transform Results
Tools : Python – Collab Notebook

datamart : gabungan dari SD Coklat, Jingga, Ungu setelah transform

ID	Std-NIS	NIS	school	fullName	email	gender	grade	POB	birth	age	caturWulan	subjects	midTest	finalTest	mean	finalScore
1	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	1	Math	95	82	88.5	A
2	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	2	Math	66	76	71.0	B
3	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	3	Math	59	84	71.5	B

Kolom baru :ID, email

Case and Solving

- Kepala Dinas menginginkan adanya standar Penilaian sebagai berikut:
  - Nilai Rata-rata 50-69 mendapat nilai C
  - Nilai Rata-rata 70-84 mendapat nilai B
  - Nilai Rata-rata 85-100 mendapat nilai A

Tambahkan kolom Nilai Mutu A,B atau C

```
# Tambahkan kolom 'Nilai Mutu' berdasarkan ketentuan
data_coklat_combined_2['Nilai Mutu'] = pd.cut(
    data_coklat_combined['Rata-Rata'],
    bins=[0, 49, 69, 84, 100],
    labels=['D', 'C', 'B', 'A'],
    right=True
)
```

inklusif

Case and Solving

Kepala Dinas ingin tiap siswa memiliki alamat email

- domain : @sekolah.edu.id
- Alamat email adalah gabungan sbb
  - Huruf awal / initial nama depan
  - Full nama belakang
  - Contoh : Antonius Binsar Tarigan → a.tarigan@sekolah.edu.id

Buat Email

```
# Fungsi untuk membuat email
def create_email(row):
    first_name_initial = row['fullName'].split()[0][0].lower()
    last_name = row['fullName'].split()[-1].lower()
    email = f"{first_name_initial}.{last_name}@sekolah.edu.id"
    return email

dataMart['email'] = dataMart.apply(create_email, axis=1)
```

Case and Solving

1. Buatlah **satu** tabel datamart yang menampung semua data SD dengan syarat sebagai berikut
   a. Standarisasi NIS
   b. Penambahan kolom email dengan format seperti diatas. Satu alamat email untuk satu siswa.
   c. Standarisasi tanggal
   d. Standarisasi jenis kelamin

ID	Std-NIS	NIS	school	fullName	email	gender	grade	POB	birth	age	caturWulan	subjects	midTest	finalTest	mean	finalScore
1	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	1	Math	95	82	88.5	A
2	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	2	Math	66	76	71.0	B
3	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	3	Math	59	84	71.5	B

Case and Solving

Lanjutan : Standarisasi NIS

```
dataMart['Std-NIS'] = dataMart['NIS'] + '-' + dataMart['school'].str.split().str[1]
dataMart
```

Standarisasi Tanggal

```
data_coklat_combined_2['Tanggal Lahir'] = pd.to_datetime(data_coklat_combined_2['YOB'].astype(str) + '-' + data_coklat_combined_2['MOB'].astype(str) + '-' + data_coklat_combined_2['DOB'].astype(str), errors='coerce')
```

Standarisasi Jenis Kelamin

```
# Mengganti nilai pada kolom 'Jenis Kelamin'
sheet1['Jenis Kelamin'] = sheet1['Jenis Kelamin'].replace({'P': 'M', 'W': 'F'})
sheet1.head()
```

Case and Solving

2. Buatlah satu table summary dari point nomor satu dengan syarat sebagai berikut

   a. Hilangkan kolom nilai UTS dan UAS per siswa per masing-masing mata pelajaran, diganti dengan Term Average yang merupakan nilai rata-rata (average) UTS + UAS
   b. Tambahkan kolom Final Score persiswa per mata pelajaran dengan indikator huruf (A,B,C)
   c. Tambahkan umur per hari ini untuk setiap siswa. Jika umur tidak bulat, maksimal tambahkan satu angka di belakang koma.

Case and Solving

ID	Std-NIS	NIS	school	fullName	email	gender	grade	POB	birth	age	caturWulan	subjects	mean	finalScore
1	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	1	Math	88.5	A
2	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	2	Math	71.0	B
3	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	3	Math	71.5	B
4	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	1	Science	73.0	B
5	Std001-Coklat	Std001	SD Coklat	Ika Tiara Alatas	i.alatas@sekolah.edu.id	F	6	Pariaman	2011-01-05	13	2	Science	83.5	B
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16196	Std300-Ungu	Std300	SD Ungu	Kemala Agusta Taslim	k.taslim@sekolah.edu.id	F	1	Langsa	2014-07-23	9	2	English	69.5	B
16197	Std300-Ungu	Std300	SD Ungu	Kemala Agusta Taslim	k.taslim@sekolah.edu.id	F	1	Langsa	2014-07-23	9	3	English	61.0	C
16198	Std300-Ungu	Std300	SD Ungu	Kemala Agusta Taslim	k.taslim@sekolah.edu.id	F	1	Langsa	2014-07-23	9	1	Religion	81.5	B
16199	Std300-Ungu	Std300	SD Ungu	Kemala Agusta Taslim	k.taslim@sekolah.edu.id	F	1	Langsa	2014-07-23	9	2	Religion	64.0	C
16200	Std300-Ungu	Std300	SD Ungu	Kemala Agusta Taslim	k.taslim@sekolah.edu.id	F	1	Langsa	2014-07-23	9	3	Religion	88.0	A

Case and Solving

Tabel Summary Soal 2

	count	unique	top	freq
Std-NIS	16200	900	Std001-Coklat	18
NIS	16200	300	Std001	54
school	16200	3	SD Coklat	5400
fullName	16200	896	Reni Prita Butarbutar	54
email	16200	758	a.julio@sekolah.edu.id	72
gender	16200	2	M	8640
POB	16182	102	Solok	324
birth	16200	742	2011-12-22	72
subjects	16200	6	Math	2700
finalScore	16200	3	B	8054

	count	mean	std	min	25%	50%	75%	max
ID	16200.0	8100.500000	4676.681516	1.0	4050.75	8100.5	12150.25	16200.0
grade	16200.0	3.500000	1.707878	1.0	2.00	3.5	5.00	6.0
age	16200.0	10.091111	2.014049	5.0	9.00	10.0	12.00	15.0
caturWulan	16200.0	2.000000	0.816522	1.0	1.00	2.0	3.00	3.0
mean	16200.0	74.965617	10.407333	50.0	67.50	75.0	82.50	100.0

Case and Solving

Syntax Summary Soal 2

Age :

```
# Hitung umur per hari ini
data_ungu_combined_2['Umur'] = (datetime.now() - data_ungu_combined_2['birth']).astype('<m8[Y]').astype(int)
```

mean :

```
# Mengisi kolom 'Rata-Rata' dengan nilai rata-rata (Mid Test + Final Test)/2
data_jingga_combined['Rata-Rata'] = (data_jingga_combined['Mid Test'] + data_jingga_combined['Final Test']) / 2
```

finalScore : Hal 11

Soal no 3

3. Buatlah satu table summary dari point nomor 2 dengan syarat sebagai berikut

Tampilkan informasi dengan detail

a. Sekolah
b. Nilai rata-rata Term Average per mata pelajaran per catur wulan per grade dalam bentuk angka dan huruf
c. Nilai rata-rata per mata pelajaran per grade dalam bentuk angka dan huruf
d. Rata-rata usia per grade per sekolah.
e. Jumlah masing-masing gender per grade per sekolah

Soal no 3

Nilai rata-rata Term Average per mata pelajaran per catur wulan per grade dalam bentuk angka dan huruf

subjects	caturWulan	grade	mean	huruf
Bahasa	1	1	76.13	B
Bahasa	1	2	74.38	B
Bahasa	1	3	74.84	B
Bahasa	1	4	74.67	B
Bahasa	1	5	74.51	B
...	...	...	...	...
Social	3	2	75.56	B
Social	3	3	75.02	B
Social	3	4	74.55	B
Social	3	5	74.98	B
Social	3	6	75.74	B

Nilai rata-rata Term Average per mata pelajaran per catur wulan per grade dalam bentuk angka dan huruf

```
average_mean_table1 = dataMart.groupby(['subjects', 'caturWulan', 'grade'])['mean'].mean().reset_index()
average_mean_table1['mean'] = average_mean_table1['mean'].round(2)

# Tambahkan kolom 'huruf' berdasarkan ketentuan
average_mean_table1['huruf'] = pd.cut(
    average_mean_table1['mean'],
    bins=[0, 49, 69, 84, 100],
    labels=['D', 'C', 'B', 'A'],
    right=True
)

average_mean_table1
```

Soal no 3

Nilai rata-rata per mata pelajaran per grade dalam bentuk angka dan huruf

	subjects	grade	mean	huruf
0	Bahasa	1	74.61	B
1	Bahasa	2	74.36	B
2	Bahasa	3	74.97	B
3	Bahasa	4	75.23	B
4	Bahasa	5	75.19	B
5	Bahasa	6	75.24	B
6	English	1	74.78	B
7	English	2	74.77	B
8	English	3	74.29	B
9	English	4	74.83	B
10	English	5	74.63	B

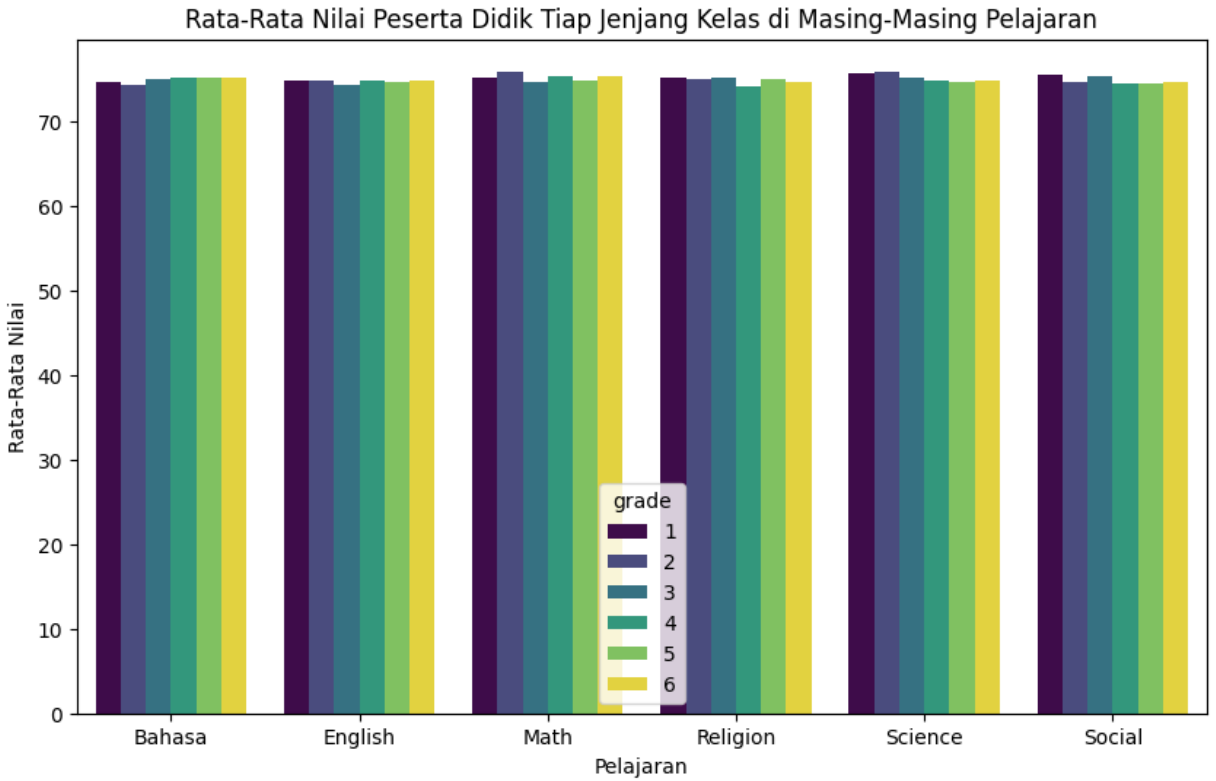
Nilai rata-rata per mata pelajaran per grade dalam bentuk angka dan huruf

```
average_mean_table2 = dataMart.groupby(['subjects', 'grade'])['mean'].mean().reset_index()
average_mean_table2['mean'] = average_mean_table2['mean'].round(2)

# Tambahkan kolom 'huruf' berdasarkan ketentuan
average_mean_table2['huruf'] = pd.cut(
    average_mean_table2['mean'],
    bins=[0, 49, 69, 84, 100],
    labels=['D', 'C', 'B', 'A'],
    right=True
)

average_mean_table2
```

Soal no 3


Rata-Rata Nilai Peserta Didik Tiap Jenjang Kelas di Masing-Masing Pelajaran
Rata-Rata Nilai
0
10
20
30
40
50
60
70
grade
1
2
3
4
5
6
Bahasa
English
Math
Religion
Science
Social
Pelajaran

Soal no 3

Rata-rata usia per grade per sekolah.

school	grade	age
SD Coklat	1	7.72
SD Coklat	2	8.60
SD Coklat	3	9.66
SD Coklat	4	10.56
SD Coklat	5	11.58
SD Coklat	6	12.64
SD Jingga	1	6.32
SD Jingga	2	8.12

Rata-rata usia per grade per sekolah.

```
average_age_table = dataMart.groupby(['school', 'grade'])['age'].mean().reset_index()
average_age_table
```

Soal no 3

e. Jumlah masing-masing gender per grade per sekolah

	school	grade	gender	count
0	SD Coklat	1	F	13
1	SD Coklat	1	M	37
2	SD Coklat	2	F	16
3	SD Coklat	2	M	34
4	SD Coklat	3	F	27
5	SD Coklat	3	M	23
6	SD Coklat	4	F	24
7	SD Coklat	4	M	26
8	SD Coklat	5	F	22
9	SD Coklat	5	M	28

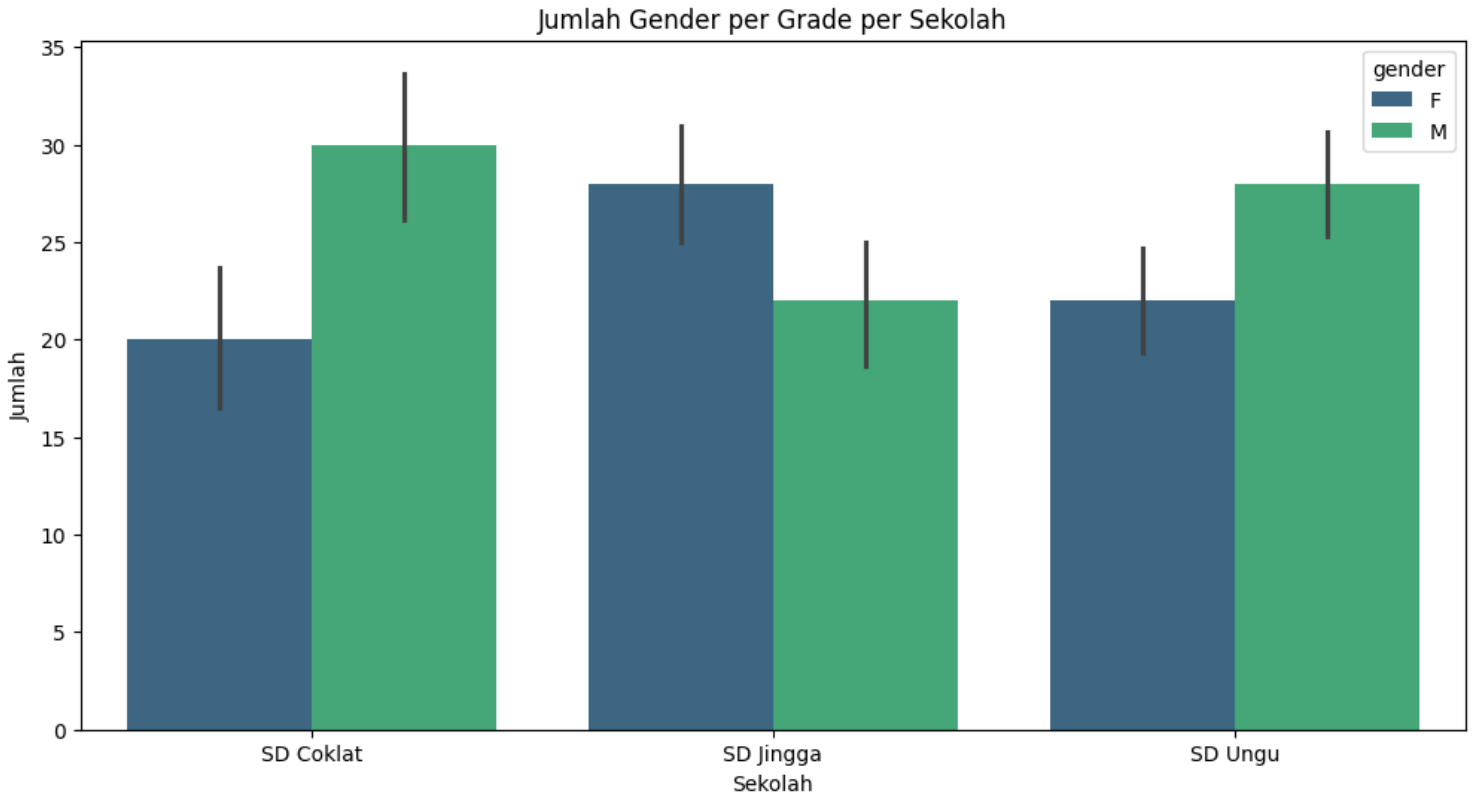
Jumlah masing-masing gender per grade per sekolah

```
# Mengambil kolom yang diperlukan
selected_columns = ['Std-NIS', 'school', 'grade', 'gender']
reduced_data = dataMart[selected_columns]

# Menghapus duplikat berdasarkan 'Std-NIS'
unique_data = reduced_data.drop_duplicates(subset=['Std-NIS'])
```

```
# Membuat tabel perhitungan jumlah gender per grade per sekolah
sum_gender_table = unique_data.groupby(['school', 'grade','gender']).size().reset_index(name='count')
sum_gender_table
```

Soal no 3


Jumlah Gender per Grade per Sekolah
gender
F
M
Jumlah
0
5
10
15
20
25
30
35
SD Coklat
SD Jingga
SD Ungu
Sekolah

Soal no 4

4. Berikan kesimpulan kepada Kepala Dinas
   a. Urutan sekolah dengan nilai rata-rata per mata pelajaran per grade dari yang tertinggi sampai ke terendah.
   b. Urutan sekolah dengan jenjang usia tertua sampai termuda di tiap grade.

Soal no 4

Urutan sekolah dengan nilai rata-rata per mata pelajaran per grade dari yang tertinggi sampai ke terendah.

subjects	grade	school	mean
Bahasa	1	SD Ungu	75.410000
Bahasa	1	SD Jingga	74.493333
Bahasa	1	SD Coklat	73.913333
Bahasa	2	SD Ungu	74.733333
Bahasa	2	SD Jingga	74.423333
...	...	...	...
Social	5	SD Ungu	73.956667

Urutan sekolah dengan nilai rata-rata per mata pelajaran per grade dari yang tertinggi sampai ke terendah.

```
# Membuat tabel perhitungan rata-rata mean per mata pelajaran per grade per sekolah
average_subjects_table = dataMart.groupby(['subjects', 'grade', 'school'])['mean'].mean().reset_index()

# Mengurutkan DataFrame berdasarkan nilai rata-rata dari yang tertinggi ke terendah
sorted_table = average_subjects_table.sort_values(by=['subjects', 'grade', 'mean'], ascending=[True, True, False])
# sorted_table['mean'] = sorted_table['mean'].round(2)

sorted_table
```

Soal no 4

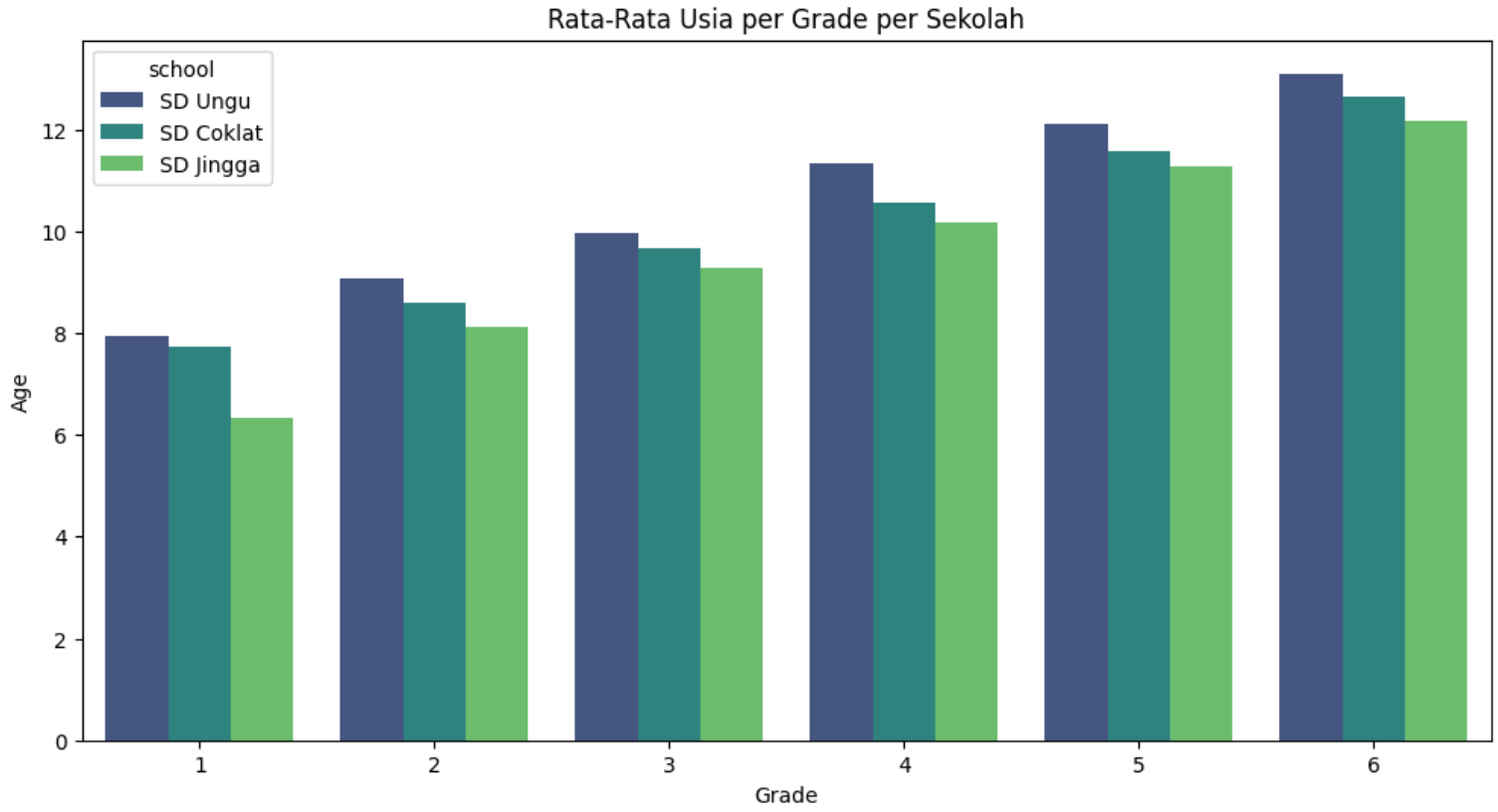
Urutan sekolah dengan jenjang usia tertua sampai termuda di tiap grade.

grade	school	age
1	SD Ungu	7.94
1	SD Coklat	7.72
1	SD Jingga	6.32
2	SD Ungu	9.08
2	SD Coklat	8.60
2	SD Jingga	8.12
3	SD Ungu	9.98
3	SD Coklat	9.66

Urutan sekolah dengan jenjang usia tertua sampai termuda di tiap grade

```
average_age_table = dataMart.groupby(['grade','school'])['age'].mean().reset_index()
average_age_sort = average_age_table.sort_values(by=['grade', 'age'], ascending=[True, False])
average_age_sort
```

Soal no 4


Rata-Rata Usia per Grade per Sekolah
school
SD Ungu
SD Coklat
SD Jingga
Age
Grade
0
2
4
6
8
10
12
1
2
3
4
5
6

Terima Kasih

NAUFAL NASHIF

https://github.com/naufalnashif/data-engineer-test-case