

Tugas 1 Data Sains dan Analisis

Menelaah Data dengan Statistik

1. Menampilkan Data

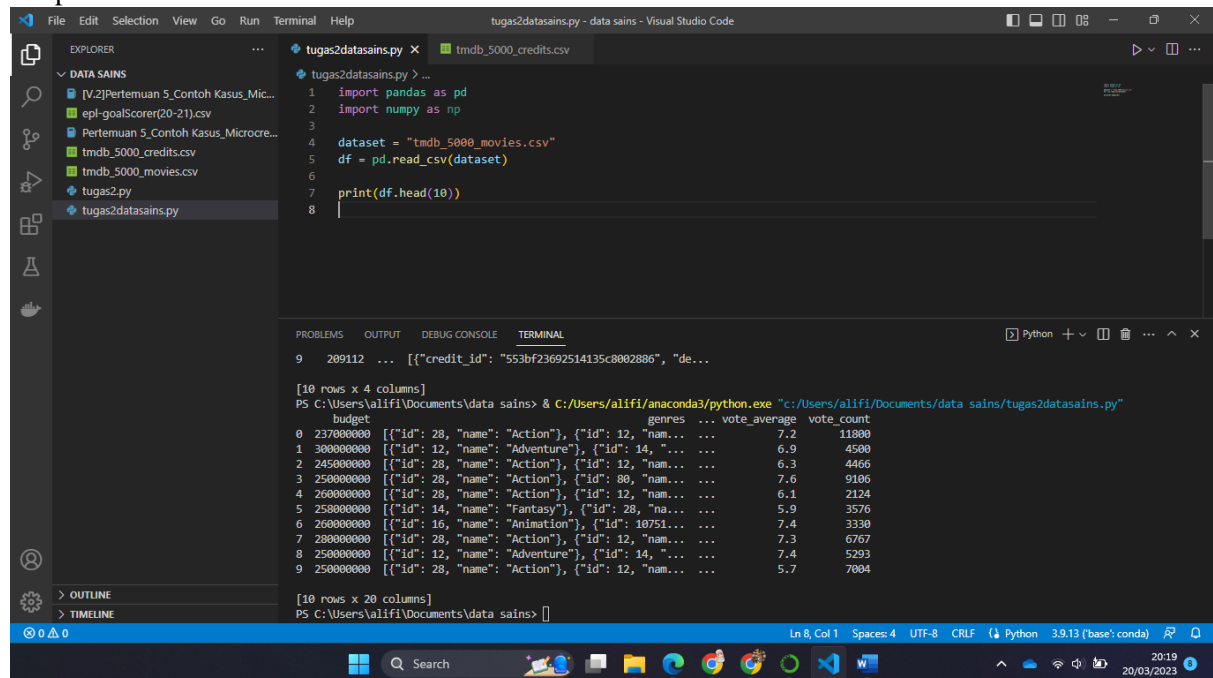
Source Code

```
import pandas as pd
import numpy as np

dataset = "tmdb_5000_movies.csv"
df = pd.read_csv(dataset)

print(df.head(10))
```

Output



The screenshot shows the Visual Studio Code interface with a file explorer on the left, a code editor in the center, and a terminal at the bottom. The code editor displays the same Python code as the Source Code block. The terminal shows the output of the script, which includes the first 10 rows of the dataset. The output is as follows:

```
[10 rows x 4 columns]
PS C:\Users\alifi\Documents\data sains> C:\Users\alifi\anaconda3\python.exe "C:\Users\alifi\Documents\data sains\tugas2datasains.py"
budget
0 237000000 [{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 7.2 11800
1 300000000 [{"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 6.9 4500
2 245000000 [{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 6.3 4466
3 250000000 [{"id": 28, "name": "Action"}, {"id": 80, "name": "Comedy"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 7.6 9106
4 260000000 [{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 6.1 2124
5 258000000 [{"id": 14, "name": "Fantasy"}, {"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 5.9 3576
6 260000000 [{"id": 16, "name": "Animation"}, {"id": 10751, "name": "Documentary"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 7.4 3330
7 280000000 [{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 7.3 6767
8 250000000 [{"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 7.4 5293
9 250000000 [{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 16, "name": "Animation"}] 5.7 7004

[10 rows x 20 columns]
PS C:\Users\alifi\Documents\data sains>
```

2. Menelaah Data

Source Code

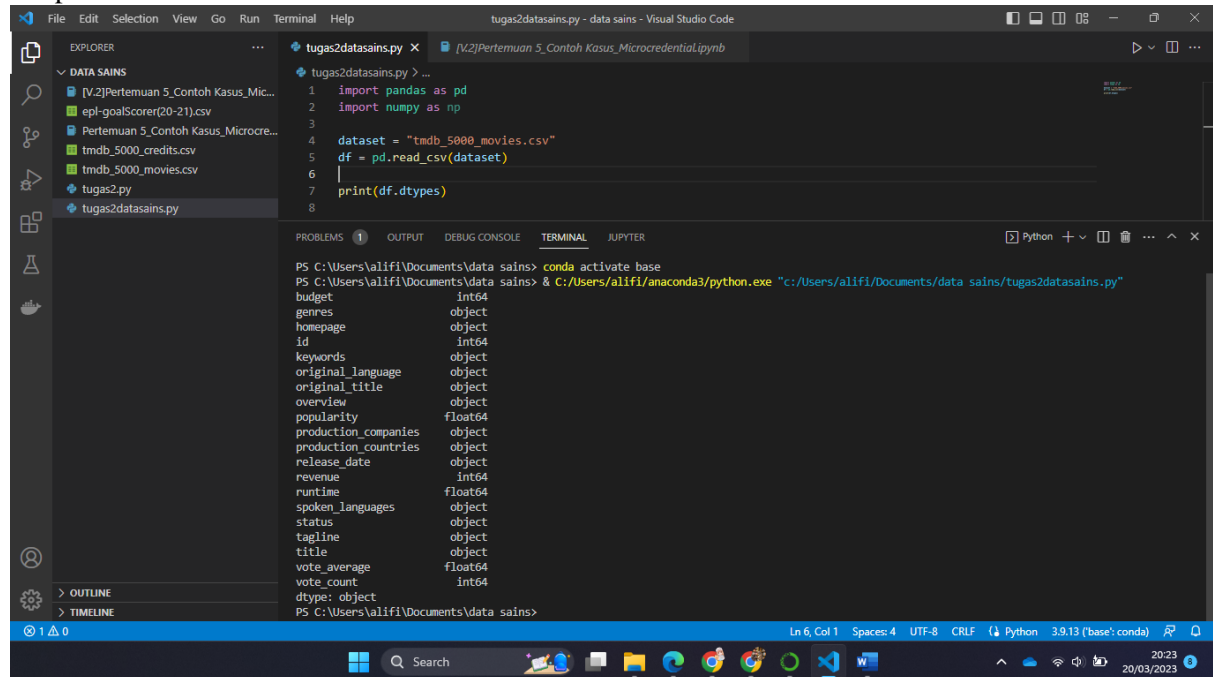
df.dtypes memungkinkan untuk melihat tipe-tipe data dari setiap kolom

```
import pandas as pd
import numpy as np

dataset = "tmdb_5000_movies.csv"
df = pd.read_csv(dataset)

print(df.dtypes)
```

Output



```
tugas2datasains.py - data sains - Visual Studio Code

EXPLORER
DATA SAINS
  [V2]Pertemuan 5_Contoh Kasus_Mic...
  epl-goalScorer(20-21).csv
  Pertemuan 5_Contoh Kasus_Microcre...
  tmdb_5000_credits.csv
  tmdb_5000_movies.csv
  tugas2.py
  tugas2datasains.py

tugas2datasains.py
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 print(df.dtypes)
8

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
Python + - [] ... ^ x

PS C:\Users\alifi\Documents\data sains> conda activate base
PS C:\Users\alifi\Documents\data sains> & C:/Users/alifi/anaconda3/python.exe "c:/Users/alifi/Documents/data sains/tugas2datasains.py"
budget int64
genres object
homepage object
id int64
keywords object
original_language object
original_title object
overview object
popularity float64
production_companies object
production_countries object
release_date object
revenue int64
runtime float64
spoken_languages object
status object
tagline object
title object
vote_average float64
vote_count int64
dtype: object
PS C:\Users\alifi\Documents\data sains>
```

Source Code

df_noid.describe() dapat menampilkan statistik dasar setiap kolom data yang bertipe numerik

```
df_noid = df.iloc[:,2:]
df_noid
#menampilkan statistik dasar setiap kolom data yang bertipe numerik
print(df_noid.describe())
```

Output

```

4      http://movies.disney.com/john-carter  49529  ...      6.1      2124
...
4798      NaN  9367  ...      6.6      238
4799      NaN  72766  ...      5.9      5
4800  http://www.hallmarkchannel.com/signedseadell...  231617  ...      7.0      6
4801      http://shanghaiacalling.com/  126186  ...      5.7      7
4802      NaN  25975  ...      6.3      16

[4803 rows x 18 columns]
PS C:\Users\alifi\Documents\data sains> & C:\Users\alifi\anaconda3\python.exe "C:\Users\alifi\Documents\data sains\tugas2datasains.py"
id popularity revenue runtime vote_average vote count
count  4803.000000  4803.000000  4.803000e+03  4801.000000  4803.000000  4803.000000
mean  57185.484281  21.492301  8.226064e+07  106.875859  6.092172  690.217989
std  88694.614033  31.816650  1.628571e+08  22.611935  1.194612  1234.585891
min  5.000000  0.000000  0.000000e+00  0.000000  0.000000  0.000000
25%  9014.500000  4.608070  0.000000e+00  94.000000  5.600000  54.000000
50%  14629.000000  12.921594  1.917000e+07  103.000000  6.200000  235.000000
75%  58610.500000  28.313505  9.291719e+07  118.000000  6.800000  737.000000
max  459488.000000  875.581305  2.787965e+09  338.000000  10.000000  13752.000000
PS C:\Users\alifi\Documents\data sains>

```

Source Code

df_noid.describe(exclude="object") dapat menampilkan statistik kolom yang bernilai non-numerik dalam kolom dengan nilai mean, minimal, maksimal, kuartil 1,2,3.

```
print(df_noid.describe(exclude="object"))
```

Output

```

PS C:\Users\alifi\Documents\data sains> C:\Users\alifi\anaconda3\Scripts\activate
PS C:\Users\alifi\Documents\data sains> conda activate base
PS C:\Users\alifi\Documents\data sains> & C:\Users\alifi\anaconda3\python.exe "C:\Users\alifi\Documents\data sains\tugas2datasains.py"
PS C:\Users\alifi\Documents\data sains> & C:\Users\alifi\anaconda3\python.exe "C:\Users\alifi\Documents\data sains\tugas2datasains.py"
id popularity revenue runtime vote_average vote count
count  4803.000000  4803.000000  4.803000e+03  4801.000000  4803.000000  4803.000000
mean  57185.484281  21.492301  8.226064e+07  106.875859  6.092172  690.217989
std  88694.614033  31.816650  1.628571e+08  22.611935  1.194612  1234.585891
min  5.000000  0.000000  0.000000e+00  0.000000  0.000000  0.000000
25%  9014.500000  4.608070  0.000000e+00  94.000000  5.600000  54.000000
50%  14629.000000  12.921594  1.917000e+07  103.000000  6.200000  235.000000
75%  58610.500000  28.313505  9.291719e+07  118.000000  6.800000  737.000000
max  459488.000000  875.581305  2.787965e+09  338.000000  10.000000  13752.000000
PS C:\Users\alifi\Documents\data sains>

```

3. Fungsi Statistik

Source Code

Fungsi mean() digunakan untuk mencari nilai mean atau rata-rata.

```
print(df_noid.mean())
```

Output

```

tugas2datasains.py
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 df_noid = df.iloc[:,1:]
8 df_noid
9 df_noid.describe()
10 df_noid.describe(exclude="object")
11
12 #Contoh fungsi statistik setiap kolom
13 print(df_noid.mean())

```

```

id          5.716548e+04
popularity  2.149238e+01
revenue     8.226864e+07
runtime     1.068759e+02
vote_average 6.092172e+00
vote_count  6.902188e+02
dtype: float64
PS C:\Users\alifi\Documents\data_sains>

```

Source Code

Fungsi sum() digunakan untuk menjumlahkan bilangan numerik berdasarkan kolom.

```
print(df_noid.sum())
```

Output

```

tugas2datasains.py
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 df_noid = df.iloc[:,1:]
8 df_noid
9 df_noid.describe()
10 df_noid.describe(exclude="object")
11
12 #Contoh fungsi statistik setiap kolom
13 #print(df_noid.mean())
14 print(df_noid.sum())
15 #print(df_noid.median())

```

```

original_language      enenenenenenenenenenenenenenenenenenen...
original_title      AvatarPirates of the Caribbean: At World's End...
popularity              103227.519725
production_companies  [{"name": "Ingenious Film Partners", "id": 289...
production_countries  [{"iso_3166_1": "US", "name": "United States o...
revenue                395097847444
runtime                913111.0
spoken_languages      [{"iso_639_1": "en", "name": "English"}, {"iso...
status                ReleasedReleasedReleasedReleasedReleasedReleas...
title                AvatarPirates of the Caribbean: At World's End...
vote_average          29260.7
vote_count            3315117
dtype: object
PS C:\Users\alifi\Documents\data_sains>

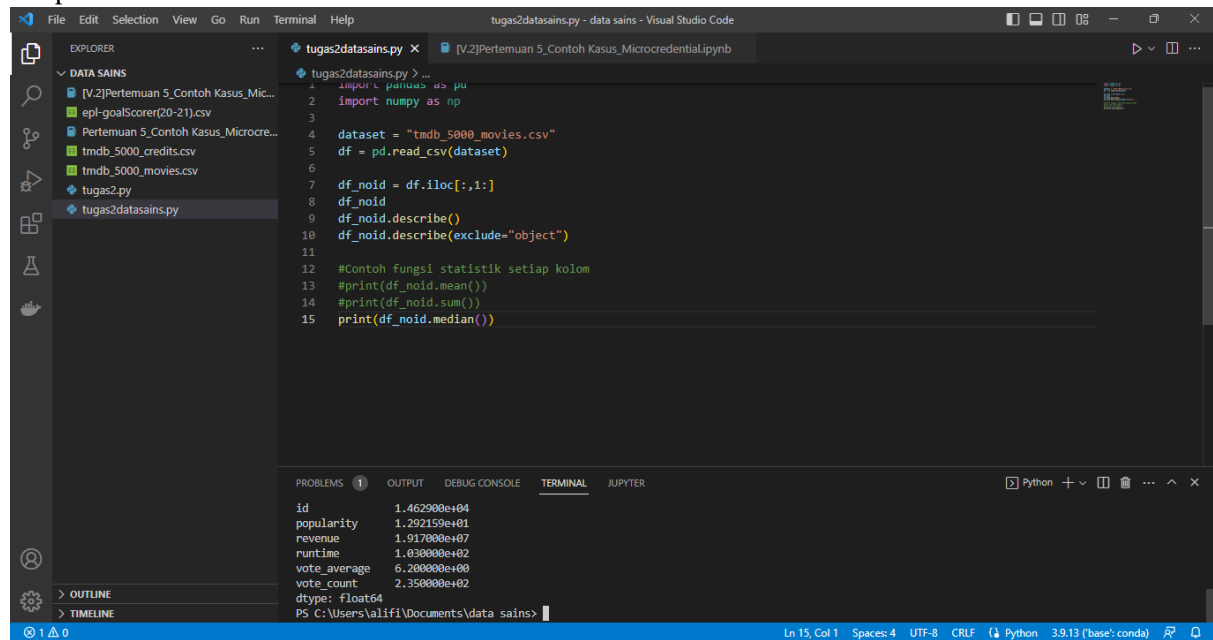
```

Source Code

Fungsi median() digunakan untuk mencari nilai tengah dari data yang bertipe numerik.

```
print(df_noid.median())
```

Output



```
tugas2datasains.py
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 df_noid = df.iloc[:,1:]
8 df_noid
9 df_noid.describe()
10 df_noid.describe(exclude="object")
11
12 #Contoh fungsi statistik setiap kolom
13 #print(df_noid.mean())
14 #print(df_noid.sum())
15 print(df_noid.median())
```

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

id	1.462900e+04
popularity	1.292159e+01
revenue	1.917000e+07
runtime	1.030000e+02
vote_average	6.200000e+00
vote_count	2.350000e+02
dtype:	float64

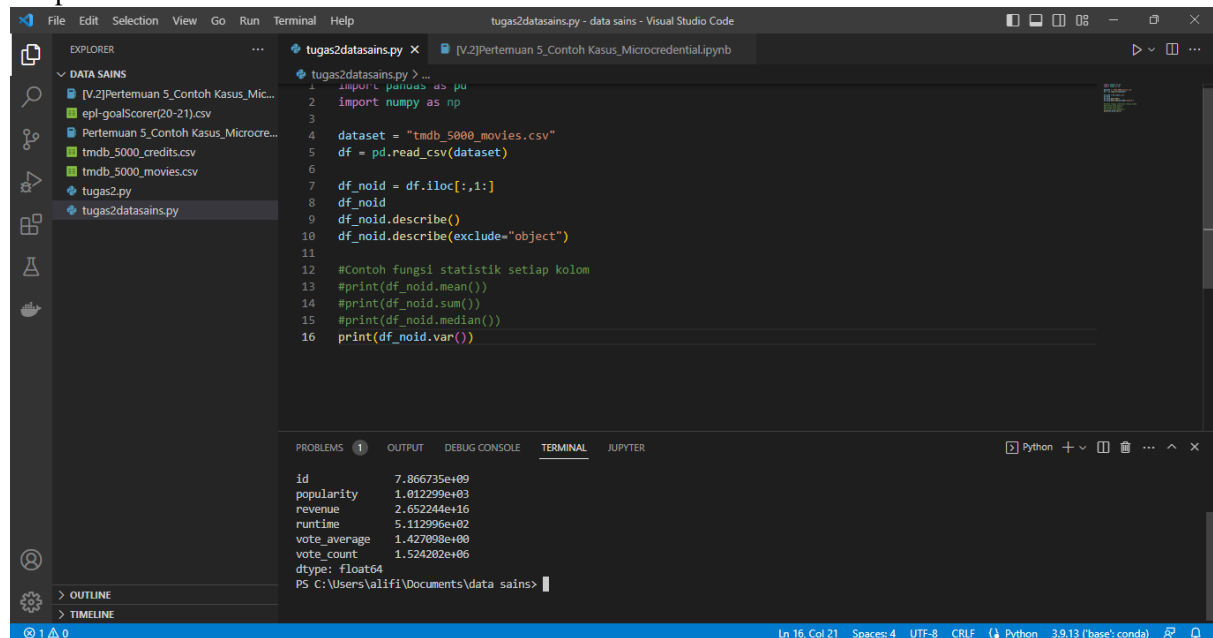
PS C:\Users\alifi\Documents\data_sains>

Source Code

Fungsi var() digunakan untuk menentukan nilai varian.

```
print(df_noid.var())
```

Output



```
tugas2datasains.py
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 df_noid = df.iloc[:,1:]
8 df_noid
9 df_noid.describe()
10 df_noid.describe(exclude="object")
11
12 #Contoh fungsi statistik setiap kolom
13 #print(df_noid.mean())
14 #print(df_noid.sum())
15 #print(df_noid.median())
16 print(df_noid.var())
```

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

id	7.866735e+09
popularity	1.012299e+03
revenue	2.652244e+16
runtime	5.112996e+02
vote_average	1.427098e+00
vote_count	1.524202e+06
dtype:	float64

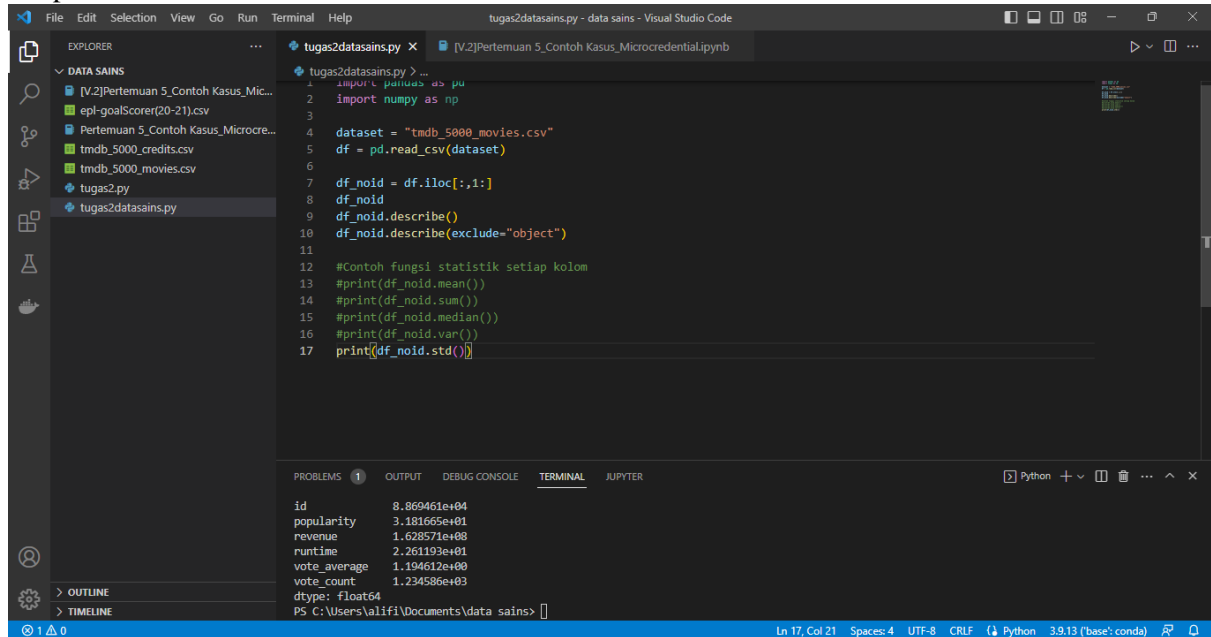
PS C:\Users\alifi\Documents\data_sains>

Source Code

Fungsi std() digunakan untuk mencari nilai standar deviasi.

```
print(df_noid.std())
```

Output



```
tugas2datasains.py
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 df_noid = df.iloc[:,1:]
8 df_noid
9 df_noid.describe()
10 df_noid.describe(exclude="object")
11
12 #Contoh fungsi statistik setiap kolom
13 #print(df_noid.mean())
14 #print(df_noid.sum())
15 #print(df_noid.median())
16 #print(df_noid.var())
17 print(df_noid.std())
```

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

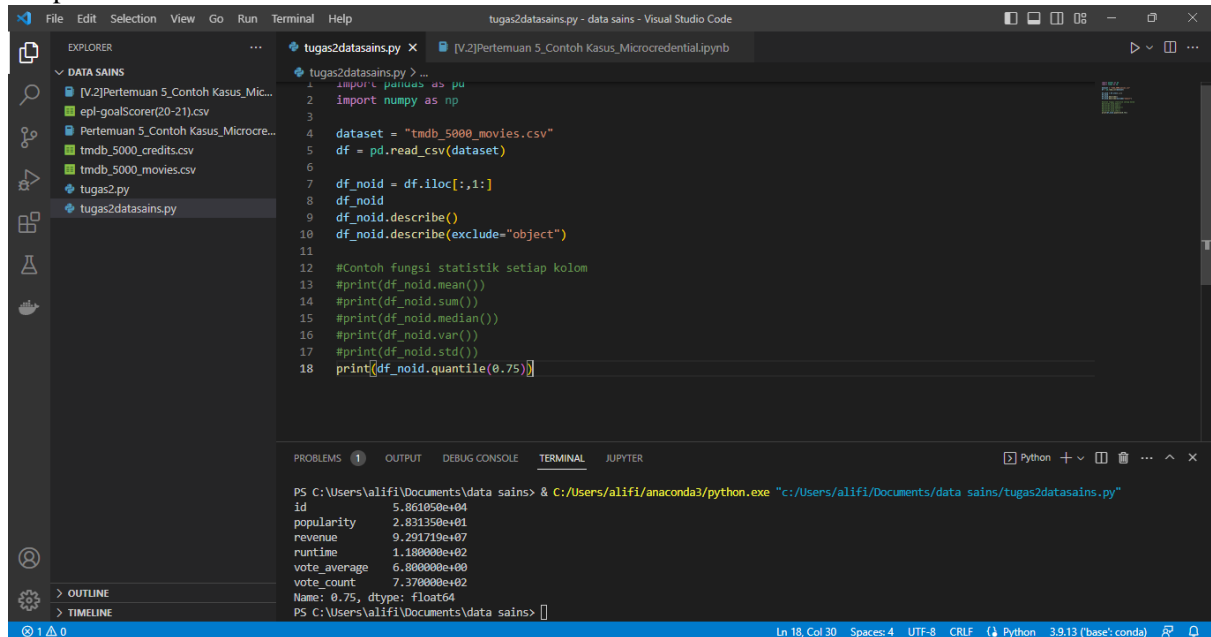
```
id 8.869461e+04
popularity 3.181665e+01
revenue 1.628571e+08
runtime 2.261193e+01
vote_average 1.194612e+00
vote_count 1.234586e+03
dtype: float64
PS C:\Users\alifi\Documents\data sains>
```

Source Code

Fungsi quartile(0.75) digunakan untuk menghitung nilai kuartil 3. Sementara fungsi kuartil 1 ialah quartile(0.25) sedangkan fungsi kuartil 2 adalah quartile(0.5).

```
print(df_noid.quantile(0.75))
```

Output



```
tugas2datasains.py
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 df_noid = df.iloc[:,1:]
8 df_noid
9 df_noid.describe()
10 df_noid.describe(exclude="object")
11
12 #Contoh fungsi statistik setiap kolom
13 #print(df_noid.mean())
14 #print(df_noid.sum())
15 #print(df_noid.median())
16 #print(df_noid.var())
17 #print(df_noid.std())
18 print(df_noid.quantile(0.75))
```

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

```
PS C:\Users\alifi\Documents\data sains> & C:\Users\alifi\anaconda3\python.exe "c:/Users/alifi/Documents/data sains/tugas2datasains.py"
id 5.861050e+04
popularity 2.831350e+01
revenue 9.291719e+07
runtime 1.180000e+02
vote_average 6.800000e+00
vote_count 7.370000e+02
Name: 0.75, dtype: float64
PS C:\Users\alifi\Documents\data sains>
```

Source Code

```
#Mencari pencilan dengan Turkey's fences (1)
q1 = df_noid.quantile(0.25)
q3 = df_noid.quantile(0.75)
iqr = q3 - q1
print(iqr)
```

Output

```
PS C:\Users\alifi\Documents\data_sains> & C:\Users\alifi\anaconda3\python.exe "c:/Users/alifi/Documents/data_sains/tugas2datasains.py"
id          4.959600e+04
popularity  2.364343e+01
revenue     9.291719e+07
runtime     2.480000e+01
vote_average 1.200000e+00
vote_count  6.830000e+02
dtype: float64
PS C:\Users\alifi\Documents\data_sains>
```

Source Code

```
#Mencari pencilan dengan Turkey's fences (2)
import warnings
warnings.filterwarnings('ignore')
# outlier filter
df_noid_align, iqr_new = df_noid.align(iqr, axis=1, copy=False,
join='outer')
outlier_filter = (df_noid < q1 - 1.5 * iqr_new) | (df_noid > q3 + 1.5 *
iqr_new)
print(outlier_filter)
```

Output

```
PS C:\Users\alifi\Documents\data_sains> & C:\Users\alifi\anaconda3\python.exe "c:/Users/alifi/Documents/data_sains/tugas2datasains.py"
genres homepage id keywords original_language original_title ... spoken_languages status tagline title vote_average vote_coun
t 0 False False False False False False ... False False False False False True
e 1 False False False False False False ... False False False False False True
e 2 False False True False False False ... False False False False False True
e 3 False False False False False False ... False False False False False True
e 4 False False False False False False ... False False False False False True
e ... ... ... ... ... ... ... ... ... ... ... ..
4798 False False False False False False ... False False False False False Fals
4799 False False False False False False ... False False False False False Fals
4800 False False True False False False ... False False False False False Fals
4801 False False False False False False ... False False False False False Fals
4802 False False False False False False ... False False False False False Fals
e
[4803 rows x 19 columns]
PS C:\Users\alifi\Documents\data_sains>
```

Source Code

```
print(df_noid[['original_title','popularity']].value_counts())
```

Output

The screenshot shows the Visual Studio Code interface with a file explorer on the left containing files like 'DATA SAINS', 'tugas2.py', and 'tugas2datasains.py'. The main editor displays the source code for 'tugas2datasains.py', which imports pandas and numpy, reads a CSV file, and prints the value counts for 'original_title' and 'popularity'. The terminal at the bottom shows the command to run the script, which outputs a table of popularity counts for various movie titles.

```
PS C:\Users\alifi\Documents\data_sains> C:/Users/alifi/anaconda3/Scripts/activate
PS C:\Users\alifi\Documents\data_sains> conda activate base
PS C:\Users\alifi\Documents\data_sains> & C:/Users/alifi/anaconda3/python.exe "c:/Users/alifi/Documents/data_sains/tugas2datasains.py"
original_title      popularity
#Horror              2.815228      1
Spring Breakers     62.554173      1
Split Second        4.857028      1
Splice              23.227877      1
Splash              21.735068      1
..                  ..
Hellboy II: The Golden Army  58.579760      1
Hellboy             47.479755      1
Hell's Angels        8.484123      1
Hell                3.629777      1
해운대              4.628525      1
Length: 4803, dtype: int64
PS C:\Users\alifi\Documents\data_sains>
```

Source Code

Menganalisis dengan metode groupby. Metode ini memungkinkan analisis dilakukan secara berkelompok berdasarkan nilai atribut yang sudah ditentukan, dalam source code menggunakan 'original_title' dan popularitasnya. Metode groupby yang digunakan ialah std() untuk mencari standar deviasi.

```
print(df.groupby('original_title')['popularity'].std())
```

Output

The screenshot shows the Visual Studio Code interface with the same file explorer. The main editor displays the source code for 'tugas2datasains.py', which now uses the groupby method to calculate the standard deviation of popularity for each original title. The terminal output shows the result, which is a Series of standard deviation values for each title, with many values being NaN.

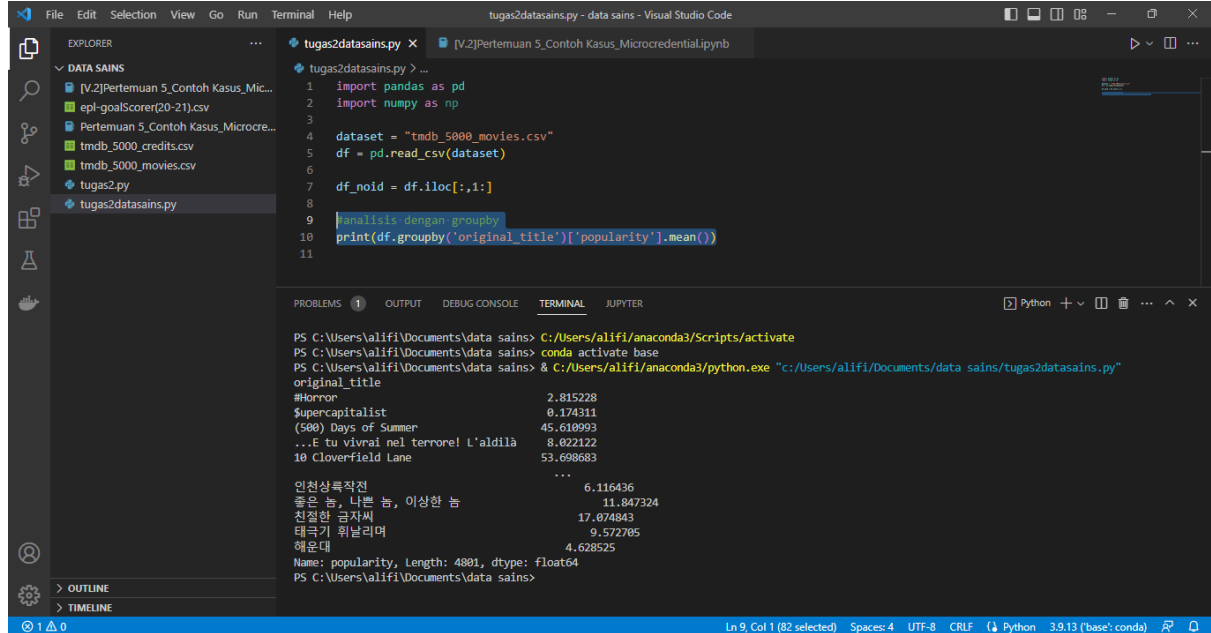
```
PS C:\Users\alifi\Documents\data_sains> C:/Users/alifi/anaconda3/Scripts/activate
PS C:\Users\alifi\Documents\data_sains> conda activate base
PS C:\Users\alifi\Documents\data_sains> & C:/Users/alifi/anaconda3/python.exe "c:/Users/alifi/Documents/data_sains/tugas2datasains.py"
original_title
#Horror      NaN
Supercapitalist      NaN
(500) Days of Summer      NaN
...E tu vivrai nel terrore! L'aldilà      NaN
10 Cloverfield Lane      NaN
..
인천상륙작전      NaN
좋은 놈, 나쁜 놈, 이상한 놈      NaN
친절한 금자씨      NaN
태극기 휘날리며      NaN
해운대      NaN
Name: popularity, Length: 4801, dtype: float64
PS C:\Users\alifi\Documents\data_sains>
```


Source Code

Menganalisis dengan metode groupby. Metode groupby yang digunakan yaitu mean() untuk mencari nilai rata-rata dari atribut 'original_title' dan popularitasnya

```
print(df.groupby('original_title')['popularity'].mean())
```

Output



```
tugas2datasains.py > ...
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 df_noid = df.iloc[:,1:]
8
9 #analisis dengan groupby
10 print(df.groupby('original_title')['popularity'].mean())
11
```

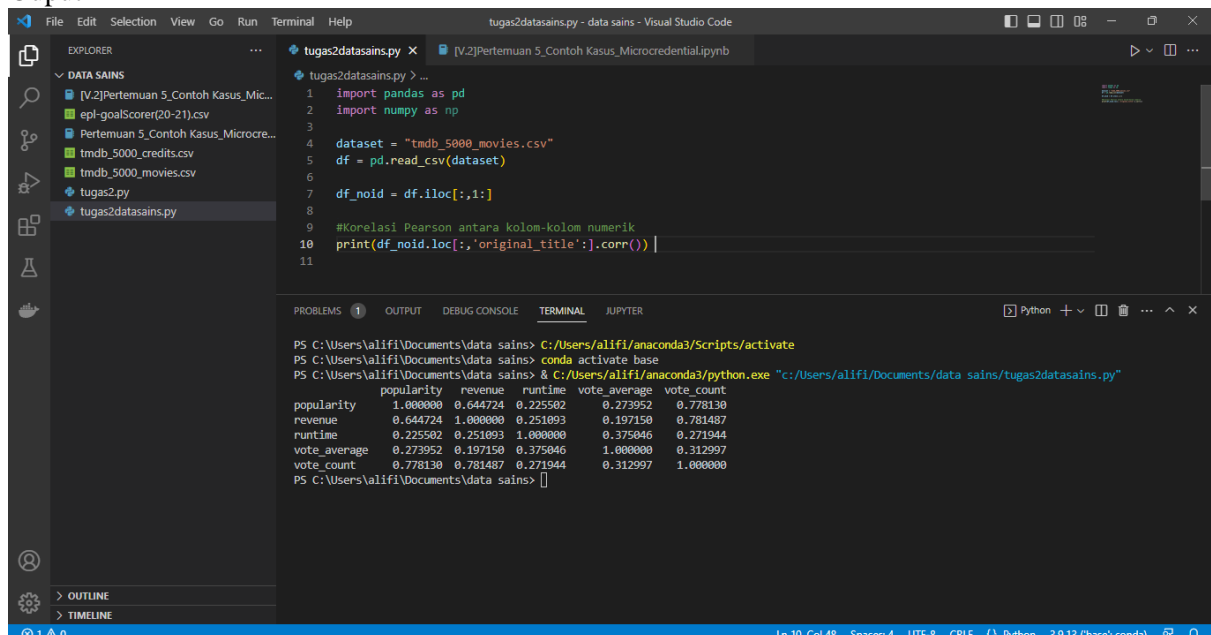
```
PS C:\Users\alifi\Documents\data_sains> C:\Users\alifi\anaconda3\Scripts\activate
PS C:\Users\alifi\Documents\data_sains> conda activate base
PS C:\Users\alifi\Documents\data_sains> & C:\Users\alifi\anaconda3\python.exe "c:/Users/alifi/Documents/data_sains/tugas2datasains.py"
original_title
#Horror                2.815228
$upercapitalist         0.174311
(500) Days of Summer   45.610993
...E tu vivrai nel terrore! L'aldilà  8.022122
10 Cloverfield Lane    53.698683
...
인천상륙작전          6.116436
좋은 놈, 나쁜 놈, 이상한 놈    11.047324
친절한 금자씨         17.074843
태극기 휘날리며       9.572705
해운대                4.628525
Name: popularity, Length: 4801, dtype: float64
PS C:\Users\alifi\Documents\data_sains>
```

Source Code

Menggunakan metode corr() untuk menghasilkan table korelasi pearson antar kolom-kolom numerik. Rentang nilainya antara -1 sampai 1. Jika bernilai -1 artinya korelasi negatif, jika bernilai 0 artinya tidak ada korelasi linier, dan jika bernilai +1 artinya korelasi positif.

```
#Korelasi Pearson antara kolom-kolom numerik
print(df_noid.loc[:, 'original_title':].corr())
```

Ouput



```
tugas2datasains.py > ...
1 import pandas as pd
2 import numpy as np
3
4 dataset = "tmdb_5000_movies.csv"
5 df = pd.read_csv(dataset)
6
7 df_noid = df.iloc[:,1:]
8
9 #Korelasi Pearson antara kolom-kolom numerik
10 print(df_noid.loc[:, 'original_title':].corr())
11
```

```
PS C:\Users\alifi\Documents\data_sains> C:\Users\alifi\anaconda3\Scripts\activate
PS C:\Users\alifi\Documents\data_sains> conda activate base
PS C:\Users\alifi\Documents\data_sains> & C:\Users\alifi\anaconda3\python.exe "c:/Users/alifi/Documents/data_sains/tugas2datasains.py"
popularity  revenue  runtime  vote_average  vote_count
popularity  1.000000  0.644724  0.225502  0.273952  0.778130
revenue      0.644724  1.000000  0.251093  0.197150  0.781487
runtime      0.225502  0.251093  1.000000  0.375046  0.271944
vote_average  0.273952  0.197150  0.375046  1.000000  0.312997
vote_count    0.778130  0.781487  0.271944  0.312997  1.000000
PS C:\Users\alifi\Documents\data_sains>
```