LAPORAN

RENCANA TUGAS MANDIRI (RTM) Ke-4

MATA KULIAH ANALISIS DATA EKSPLORATIF (A)

"Analisis Korelasi, Time Series Analysis, Hypothesis Testing dan Regresi"

DISUSUN OLEH:

Muhammad Aryasatya N (22083010085)

DOSEN PENGAMPU:

Tresna Maulana Fahrudin S.ST., M.T. (NIP. 199305012022031007)

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAWA TIMUR

2023

Tugas ini bertujuan agar mahasiswa mampu untuk melakukan langkah demi Langkah menyelesaikan permasalahan data harian COVID-19 sebagai berikut:

1. **Time Series Analysis**

a) Visualisasi data selama rentang beberapa tahun

Import library dan load dataset

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
covid19 = pd.read_csv("covid_19_indonesia_time_series_all.csv", parse_dates=['Date'], index_col='Date')
covid19
```

Output kode diatas seperti berikut:

Date	Location ISO Code	Location	New Cases	New Deaths	New Recovered	New Active Cases	Total Cases	Total Deaths	Total Recovered	Total Active Cases	...	Latitude	New Cases per Million	Total Cases per Million	New Deaths per Million	Total Deaths per Million	Total Deaths per 100rb
2020-03-01	ID-JK	DKI Jakarta	2	0	0	2	39	20	75	-56	...	-6.204699	0.18	3.60	0.0	1.84	0.18
2020-03-02	ID-JK	DKI Jakarta	2	0	0	2	41	20	75	-54	...	-6.204699	0.18	3.78	0.0	1.84	0.18
2020-03-02	IDN	Indonesia	2	0	0	2	2	0	0	2	...	-0.789275	0.01	0.01	0.0	0.00	0.00
2020-03-02	ID-RI	Riau	1	0	0	1	1	0	1	0	...	0.511648	0.16	0.16	0.0	0.00	0.00
2020-03-03	ID-JK	DKI Jakarta	2	0	0	2	43	20	75	-52	...	-6.204699	0.18	3.96	0.0	1.84	0.18
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2022-09-15	ID-SA	Sulawesi Utara	37	0	0	37	52770	1213	50997	560	...	1.259638	14.01	19974.38	0.0	459.14	45.91
2022-09-15	ID-SB	Sumatera Barat	13	0	3	10	104640	2371	102066	203	...	-0.850253	2.36	18959.11	0.0	429.59	42.96
2022-09-15	ID-SS	Sumatera Selatan	16	0	1	15	82198	3376	78510	312	...	-3.216212	1.95	10002.74	0.0	410.83	41.08
2022-09-15	ID-SU	Sumatera Utara	50	0	5	45	158866	3288	154924	654	...	2.191894	3.36	10680.15	0.0	221.04	22.10
2022-09-16	IDN	Indonesia	2358	27	2997	-666	6405044	157876	6218708	28460	...	-0.789275	8.89	24153.07	0.1	595.34	59.53

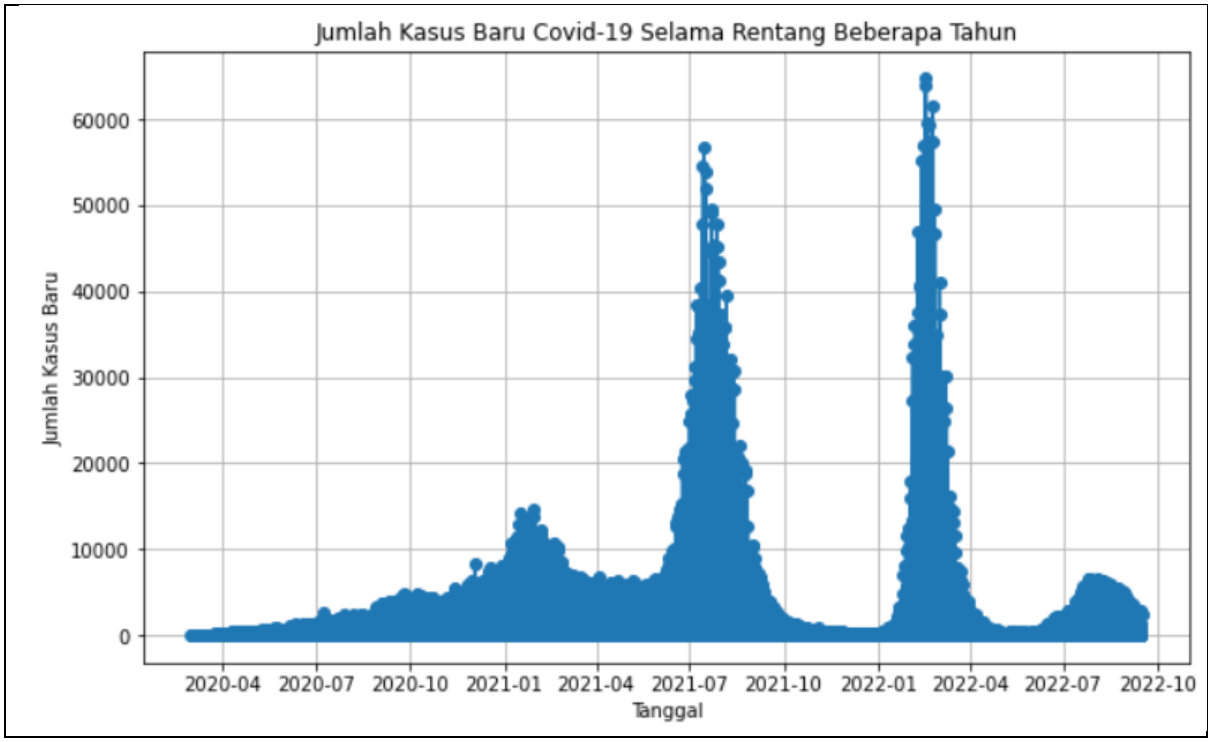
31822 rows × 37 columns

Dataset tersebut merekam perkembangan kasus COVID-19 di Indonesia dari Maret 2020 hingga September 2022. Setiap entri mencatat jumlah kasus baru, kematian, kesembuhan, dan kasus aktif pada berbagai lokasi di Indonesia. Data juga mencakup informasi geografis dan statistik seperti jumlah kasus per juta populasi, persentase kematian dan kesembuhan, serta faktor pertumbuhan kasus baru dan kematian. Ini memungkinkan analisis tren penyebaran dan dampak COVID-19 di berbagai wilayah Indonesia selama periode yang tercatat.

Selanjutnya, dilakukan visualisasi data selam rentang beberapa tahun

```
plt.figure(figsize=(10, 6))
plt.plot(covid19.index, covid19['New Cases'], marker='o', linestyle='-')
plt.title('Jumlah Kasus Baru Covid-19 Selama Rentang Beberapa Tahun')
plt.xlabel('Tanggal')
plt.ylabel('Jumlah Kasus Baru')
plt.grid(True)
plt.show()
```

Output dari kode diatas:



Grafik diatas memvisualisasikan jumlah kasus baru COVID-19 di Indonesia selama beberapa tahun dengan menggunakan rentang data yang tercatat. Dalam grafik, sumbu horizontal mewakili rentang waktu dari Maret 2020 hingga September 2022, sementara sumbu vertical menunjukkan jumlah kasus baru yang dilaporkan. Titik-titik yang terhubung oleh garis menunjukkan tren pertumbuhan kasus seiring berjalannya waktu. Dengan visualisasi ini, dapat diamati bagaimana jumlah kasus baru telah berubah dari waktu ke waktu selama periode yang tercatat.

b) Pengelompokan (Grouping) data time series berdasarkan bulan

```
month_grouped = covid19.groupby(covid19.index.month).sum()
month_grouped = month_grouped.rename_axis('Month')

month_grouped
```

Kode diatas mengelompokkan data COVID-19 berdasarkan bulan dan menjumlahkan nilai-nilai setiap bulan. Hasilnya adalah DataFrame yang menampilkan data yang terakumulasi pada setiap bulan selama periode yang diamati. Output kode:

Month	New Cases	New Deaths	New Recovered	New Active Cases	Total Cases	Total Deaths	Total Recovered	Total Active Cases	City or Regency	Total Regencies	...	Population Density	Longitude	Latitude	
1	847513	15744	571019	260750	321473186	10571546	301441501	9460139	0.0	51584	...	1577281.24	246791.477360	-5897.926257	
2	2934624	20411	1981968	932245	344142141	10044274	304084415	30013452	0.0	46592	...	1424641.12	222908.431164	-5327.159200	1
3	1251927	23182	2186009	-957264	453589225	11882295	414874225	26832705	0.0	68852	...	2247149.62	292674.107003	-7273.990433	
4	398500	12821	610473	-224794	458064519	11988742	436229590	9846187	0.0	74667	...	2287769.32	356171.570252	-8498.231240	
5	356933	12521	331215	13197	484455986	12778958	464682998	6994030	0.0	77376	...	2365921.86	370187.216040	-8846.889385	
6	843214	19299	502431	321484	484144293	12794087	461288560	10061646	0.0	74880	...	2289601.80	358245.692942	-8561.505857	
7	2799986	76279	2038690	685017	557957098	14554218	511510821	31892059	0.0	77376	...	2365921.86	370187.216040	-8846.889385	1
8	1797902	83361	2397437	-682896	636238205	17450650	590309807	28477748	0.0	77376	...	2365921.86	370187.216040	-8846.889385	1
9	560379	24777	843445	-307843	462074621	13811549	438258464	10004608	0.0	62805	...	1907942.03	298410.277701	-7133.602745	
10	305804	9357	356766	-60319	284425389	9648668	269490192	5286529	0.0	51584	...	1577281.24	246791.477360	-5897.926257	
11	282650	7253	260117	15280	283483506	9571846	269840567	4071093	0.0	49920	...	1526401.20	238830.461962	-5707.670571	
12	422921	10690	343691	68540	303969658	10162744	287787106	6019808	0.0	51584	...	1577281.24	246791.477360	-5897.926257	

12 rows × 26 columns

c) Pengelompokan (Grouping) data time series berdasarkan hari

```
month_grouped = covid19.groupby(covid19.index.day).sum()
month_grouped
```

Kode tersebut mengelompokkan data COVID-19 berdasarkan hari dalam sebulan dan menjumlahkan nilai-nilai yang tercatat pada setiap hari. Hasilnya adalah DataFrame yang menampilkan total kasus baru, kematian, kesembuhan, dan kasus aktif yang terakumulasi pada setiap hari dalam rentang waktu yang diamati. Output kode:

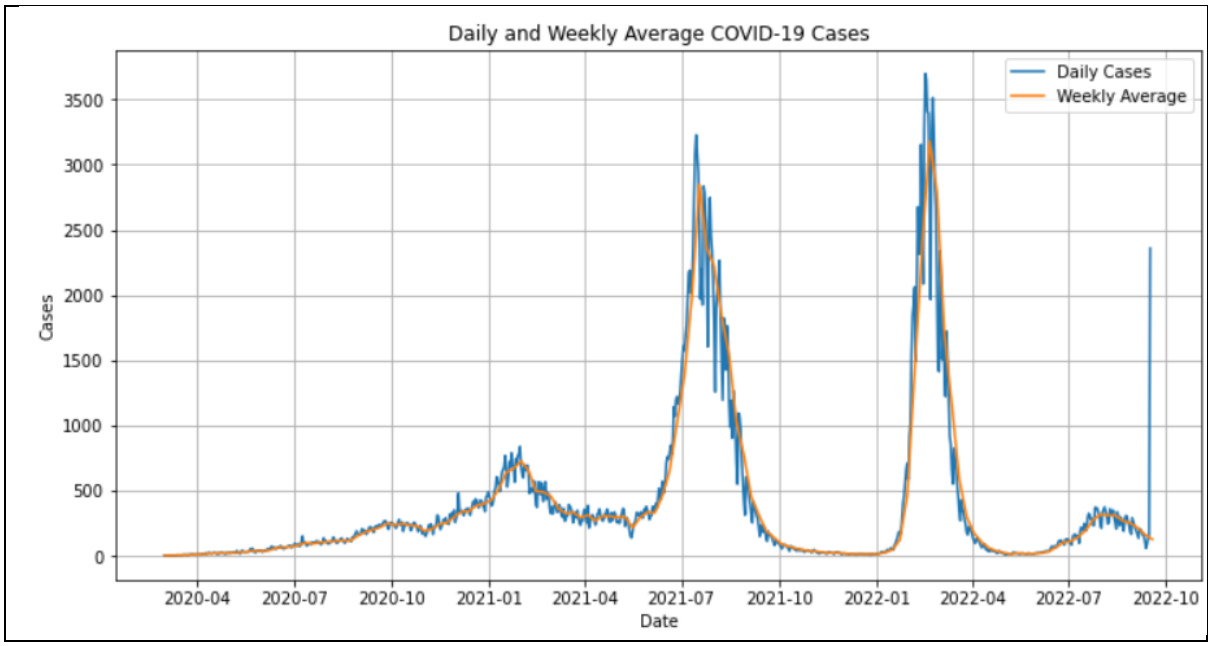
Date	New Cases	New Deaths	New Recovered	New Active Cases	Total Cases	Total Deaths	Total Recovered	Total Active Cases	City or Regency	Total Regencies	...	Population Density	Longitude	Latitude	New Case pe Millio
1	365185	10203	389722	-34740	166641557	4772768	155978357	5890432	0.0	24935	...	779318.96	119278.098770	-2852.044782	26773.2
2	381198	10538	385985	-15325	167022755	4783306	156364343	5875106	0.0	25361	...	779527.10	119493.825206	-2852.322409	27620.7
3	425885	9720	377204	38961	167448640	4793026	156741607	5914007	0.0	25379	...	780803.65	119601.428914	-2859.242841	31164.0
4	415384	10722	390330	14332	167864024	4803748	157131937	5928339	0.0	25379	...	780803.65	119601.428914	-2859.242841	30204.7
5	429385	10007	404746	14632	168293409	4813755	157536683	5942971	0.0	25379	...	780803.65	119601.428914	-2859.242841	30623.8
6	436006	10241	438112	-12347	168729415	4824001	157974905	5930509	0.0	25383	...	781913.29	119707.537919	-2865.699578	31421.0
7	407955	10843	408450	-11338	169137370	4834844	158383355	5919171	0.0	25383	...	781913.29	119707.537919	-2865.699578	28683.9
8	441681	10961	466078	-35358	169579051	4845814	158849444	5883793	0.0	25412	...	783021.93	119817.739034	-2872.958675	30734.4
9	439607	10037	415263	14307	170018658	4855851	159264707	5898100	0.0	25433	...	783133.00	119939.331305	-2881.640880	30315.7
10	431179	10727	426779	-6327	170449837	4866578	159691486	5891773	0.0	25453	...	783307.11	120183.777674	-2885.078566	30793.7
11	410299	10177	436248	-36126	170860136	4876755	160127737	5855644	0.0	25461	...	784036.54	120298.909387	-2893.448038	29529.0
12	442050	10119	448126	-16195	171302186	4886874	160575863	5839449	0.0	25461	...	784036.54	120298.909387	-2893.448038	29481.0
13	442880	9517	401444	31919	171745066	4896391	160977307	5871368	0.0	25461	...	784036.54	120298.909387	-2893.448038	30174.4
14	422623	9850	376614	36159	172167689	4906241	161353924	5907524	0.0	25468	...	784064.06	120415.377828	-2892.994180	28571.4
15	464555	9591	379838	75126	172602692	4914951	161705128	5982613	0.0	25457	...	783866.25	120173.656750	-2891.219436	31029.4

d) Visualisasi Bersama rata-rata mingguan kasus COVID-19 dan harian

```
daily_grouped = covid19.resample('D').mean()
week_grouped = covid19.resample('W').mean()

plt.figure(figsize=(12, 6))

plt.plot(daily_grouped.index, daily_grouped['New Cases'], label='Daily Cases')
plt.plot(week_grouped.index, week_grouped['  6New Cases'], label='Weekly Average')

plt.title('Daily and Weekly Average COVID-19 Cases')
plt.xlabel('Date')
plt.ylabel('Cases')
plt.legend()
plt.grid()
plt.show()
```

Kode ini menghasilkan visualisasi yang membandingkan kasus harian dengan rata-rata kasus mingguan COVID-19. Dengan menggunakan resampling ('D' untuk harian dan 'W' untuk mingguan), data yang awalnya tersedia setiap hari diubah menjadi rata-rata harian dan mingguan. Grafik menampilkan garis untuk kasus harian serta rata-rata mingguan selama rentang waktu yang diamati. Ini memungkinkan pengamat untuk melihat tren jangka pendek (hari demi hari) dan membandingkannya dengan tren jangka menengah (mingguan) dari kasus

COVID-19. Visualisasi ini dapat memberikan pemahaman yang lebih baik tentang fluktuasi sehari-hari serta tren umum dari waktu ke waktu. Output kode:



Grafik ini menggunakan dua garis, yaitu garis untuk kasus harian dan garis untuk kasus mingguan. Garis kasus harian menunjukkan jumlah kasus baru COVID-19 yang dilaporkan setiap hari, sedangkan garis kasus mingguan menunjukkan rata-rata jumlah kasus baru COVID-19 yang dilaporkan setiap minggu.

2. Hypotesis Testing

Import library yang diperlukan

```
from scipy import stats
import scipy.stats as st
import numpy as np
```

```
from scipy.stats import ttest_1samp

rata_rata_harian = covid19.resample('D').mean()['New Cases'].mean()
statistic, p_value = ttest_1samp(covid19['New Cases'], popmean=rata_rata_harian)
print(f"Nilai p-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print(f"Hipotesis nol (H0) ditolak. Rata-rata lebih besar dari {rata_rata_harian}.")
else:
    print(f"Hipotesis nol (H0) diterima. Tidak cukup bukti untuk menyatakan rata-rata lebih besar dari {rata_rata_harian}")
```

Kode diatas melakukan uji statistik untuk memeriksa apakah jumlah rata-rata COVID-19 secara signifikan lebih besar dari rata-rata harian yang diamati selama periode nilai tertentu. Dengan menggunakan uji t dan nilai p-value, kita dapat menentukan apakah terdapat cukup bukti statistik untuk menolak hipotesis bahwa rata-rata kasus baru sama dengan rata-rata harian yang teramati. Jika p-value kurang dari 0.05, maka kita menolak hipotesis nol dan menyimpulkan bahwa rata-rata kasus baru secara signifikan lebih besar dari rata-rata harian yang diamati. Di dapat output kode diatas seperti berikut:

```
Nilai p-value: 0.61990991996027
Hipotesis nol (H0) diterima. Tidak cukup bukti untuk menyatakan rata-rata lebih besar dari 395.8591735538904.
```

3. Analisis Korelasi

```python
X = covid19['Total Cases'].head(100)
Y = covid19['Total Deaths'].head(100)

Y = Y.replace(0, np.nan)
Y = Y.dropna()

n = 100

# X total
sum_x = X.sum()
sum_x_square = sum_x**2

# Xi^2
xi_square = X**2
sum_xi_square = xi_square.sum()

# Y total
sum_y = Y.sum()
sum_y_square = sum_y**2

# Yi^2
yi_square = Y**2
sum_yi_square = yi_square.sum()

# Xi * Yi
xiyi = X*Y
sum_xiyi = xiyi.sum()

p_kiri = (n*sum_xi_square) - sum_x_square
p_kanan = (n*sum_yi_square) - sum_y_square

# hitung koefisien korelasi
r = ((n*sum_xiyi) - (sum_x)*(sum_y))/((np.sqrt(p_kiri))*(np.sqrt(p_kanan)))
r
```

Kode diatas Menghitung koefisien korelasi antara total kasus dengan total kematian COVID-19 menggunakan metode Pearson. Menggunakan rumus statistik, kode ini mengolah data dari 100 entri pertama untuk kedua variabel tersebut. Prosesnya melibatkan perhitungan jumlah total, kuadrat dari setiap variabel, dan hasil perkalian silang antar keduanya. Koefisien korelasi Pearson dihitung berdasarkan rumus yang menggunakan nilai-nilai tersebut untuk mengukur seberapa eratnya hubungan linier antara total kasus dan total kematian, dengan nilai r yang berkisar antara -1 dan 1, menunjukkan arah dan kekuatan hubungan linier antara kedua variabel tersebut.

```
2.560461612200714
```

Diperoleh hasil perhitungan seperti output diatas.

4. Regresi dan Evaluasi

```
xbar = X.mean()
ybar = Y.mean()

# b1/slope
b1 = (sum_xiyi - (n*xbar*ybar))/(sum_xi_square - (n*np.sqrt(xbar)))

# b0/y-intercept
b0 = ybar - b1*xbar

# Model regresi
ytopi = b0 + b1*X
print('Model regresi = ',b0,' + ',b1,'*X')
```

```
Model regresi =  0.3973114618070692  +  0.5975854201466776 *X
```

Kode ini digunakan untuk membuat model regresi linear antara dua variabel, dalam hal ini total kasus COVID-19 dan total kematian akibat COVID-19. Dengan menggunakan metode metode Ordinary Least Squares (OLS), rumusnya menghitung kemiringan (slope) dan perpotongan sumbu y (y-intercept) untuk garis regresi. Variabel b1 menghitung kemiringan garis regresi, sedangkan b0 menghitung perpotongan sumbu y. Hasilnya adalah model regresi linier yang dapat digunakan untuk memprediksi nilai total kematian berdasarkan total kasus COVID-19. Dalam konteks ini, kode ini menyediakan perkiraan garis regresi yang paling dekat dengan pola hubungan antara kedua variabel tersebut.

```
# Evaluasi model regresi
absolute = abs((Y-ytopi)/Y)

MAPE = (1/n) * (absolute.sum()) * 100/100
MAPE
```

```
105.49417696516059
```

Kode diatas digunakan untuk Menghitung nilai MAPE untuk mengevaluasi regresi. MAPE (Mean Absolute Percentage Error) digunakan untuk mengevaluasi seberapa akurat model regresi dalam memprediksi nilai aktual dari suatu variabel. Dalam konteks ini, kode tersebut menghitung MAPE untuk mengevaluasi seberapa akurat model regresi dalam memprediksi total kematian COVID-19 berdasarkan total kasus COVID-19. MAPE mengukur kesalahan prediksi dalam bentuk persentase dari nilai aktual, menghitung rata-rata dari persentase kesalahan absolut dari setiap prediksi terhadap nilai aktualnya. Semakin rendah nilai MAPE, semakin akurat model regresi dalam memprediksi nilai aktual.