

Big Data & Predictive Analytics Final Project

# Tingkat Kunjungan Wisatawan ke Destinasi di Indonesia

Dosen Pengampu:  
Mulia Sulistiyono, M.Kom

Kelompok 6

Anggota:

1. Bakri Ahmad Ridhwan, 23.21.1576
2. Choirul Affan Adi Putra, 23.21.1572
3. Dhimas Adhiyaksa Davariza Yudhaswara, 23.21.1573

Program Studi Informatika  
Fakultas Ilmu Komputer  
Universitas Amikom Yogyakarta  
2024

## Daftar Isi

<b>1. Latar belakang</b>	<b>3</b>
<b>2. Metode</b>	<b>4</b>
2.1. Alur final project	5
2.2. Dataset	8
2.3. EDA	9
<b>3. Eksperimen</b>	<b>11</b>
<b>4. Hasil dan Evaluasi</b>	<b>18</b>
<b>3. Kesimpulan dan Kontribusi</b>	<b>20</b>
a. Kesimpulan	20
b. Kontribusi	20
<b>4. Lampiran</b>	<b>21</b>

## 1. Latar belakang

Dataset wisatawan ini mengumpulkan informasi tentang jumlah wisatawan asing yang berkunjung ke Indonesia sepanjang tahun 2024, dengan data bulanan dari Januari hingga Mei untuk beberapa negara. Data ini dikumpulkan untuk memantau tren kunjungan wisatawan, yang merupakan indikator penting dalam sektor pariwisata. Informasi ini berguna untuk mengidentifikasi pola kunjungan wisatawan dari berbagai negara dan membantu dalam pengambilan keputusan terkait strategi pemasaran dan pengembangan pariwisata.

Analisis dataset ini penting karena dapat memberikan wawasan mendalam tentang pola dan tren kunjungan wisatawan ke Indonesia. Dengan memahami distribusi jumlah wisatawan per bulan dan asal negara mereka, pemerintah dan pemangku kepentingan di sektor pariwisata dapat mengidentifikasi periode puncak dan rendah kunjungan, serta negara-negara yang paling banyak menyumbang wisatawan. Analisis ini juga dapat membantu dalam mengidentifikasi faktor-faktor yang mempengaruhi jumlah kunjungan, seperti kondisi ekonomi global, kebijakan visa, dan kampanye promosi. Dengan demikian, analisis dataset ini memungkinkan perencanaan yang lebih efektif dan strategi yang lebih tepat sasaran.

Tujuan utama dari penelitian ini adalah untuk menganalisis data kunjungan wisatawan asing ke Indonesia guna mengidentifikasi tren dan pola kunjungan. Penelitian ini bertujuan untuk menyediakan informasi yang dapat digunakan untuk meningkatkan strategi pemasaran dan pengembangan pariwisata, serta mendukung pengambilan keputusan berbasis data oleh pemerintah dan pemangku kepentingan lainnya. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi periode puncak dan rendah kunjungan, serta memahami variabilitas jumlah wisatawan dari berbagai negara sepanjang tahun. Dengan wawasan ini, diharapkan dapat diimplementasikan kebijakan dan strategi yang lebih efektif untuk meningkatkan jumlah kunjungan wisatawan ke Indonesia dan memperkuat sektor pariwisata nasional.

## 2. Metode

- Metode Analisis Deskriptif (Statistik Deskriptif dan Visualisasi Data)  
Metode analisis deskriptif adalah pendekatan yang digunakan untuk menggambarkan, meringkas, dan memahami karakteristik data tanpa membuat kesimpulan atau prediksi tentang populasi yang lebih luas. Analisis ini memberikan gambaran umum tentang data melalui statistik deskriptif dan visualisasi.
- Metode Analisis Regresi (Regresi Linier Sederhana dan Regresi Linier Berganda)  
Metode analisis regresi adalah teknik statistik yang digunakan untuk memodelkan dan menganalisis hubungan antara variabel dependen (variabel yang ingin diprediksi) dan satu atau lebih variabel independen (variabel prediktor). Analisis ini membantu dalam memahami seberapa besar pengaruh variabel independen terhadap variabel dependen dan membuat prediksi berdasarkan model yang terbentuk.

## 2.1. Alur final project

### 1. Data Collection

- Tujuan: Mengumpulkan data jumlah wisatawan asing yang berkunjung ke Indonesia dari beberapa negara selama periode tertentu.
- Proses:
  - Mengidentifikasi sumber data yang dapat diandalkan, seperti laporan kedatangan dari otoritas imigrasi atau instansi pemerintah terkait.
  - Mengumpulkan data bulanan dari Januari hingga Mei 2024 yang mencakup jumlah wisatawan dari berbagai negara.
  - Memastikan data yang dikumpulkan lengkap dan akurat untuk analisis lebih lanjut.

### 2. EDA dan Visualisasi data

- Tujuan: Memahami distribusi, tren, dan pola dalam data, serta mengidentifikasi outliers dan anomali.
- Proses:
  - Statistik Deskriptif: Menghitung statistik deskriptif seperti mean, median, mode, dan standar deviasi untuk setiap bulan dan negara.
  - Visualisasi Data:
    - Membuat grafik batang untuk menunjukkan jumlah wisatawan per bulan dari beberapa negara.
    - Membuat grafik garis untuk menunjukkan tren kunjungan wisatawan dari negara tertentu sepanjang tahun.
    - Membuat box plot untuk menunjukkan distribusi jumlah wisatawan per bulan dan mengidentifikasi outliers.
  - Insight: Mengidentifikasi bulan dengan kunjungan tertinggi dan terendah, negara asal wisatawan terbanyak, dan pola musiman dalam data.

### 3. Analisis Korelasi

- Tujuan: Menentukan hubungan antara jumlah wisatawan dari berbagai negara atau antara jumlah wisatawan dengan variabel lain (jika ada).
- Proses:
  - Menghitung koefisien korelasi Pearson untuk mengetahui kekuatan dan arah hubungan antara variabel.
  - Menggunakan heatmap untuk memvisualisasikan matriks korelasi.
  - Insight: Menentukan apakah ada hubungan yang signifikan antara kunjungan wisatawan dari satu negara dengan negara lainnya atau dengan faktor-faktor lain.

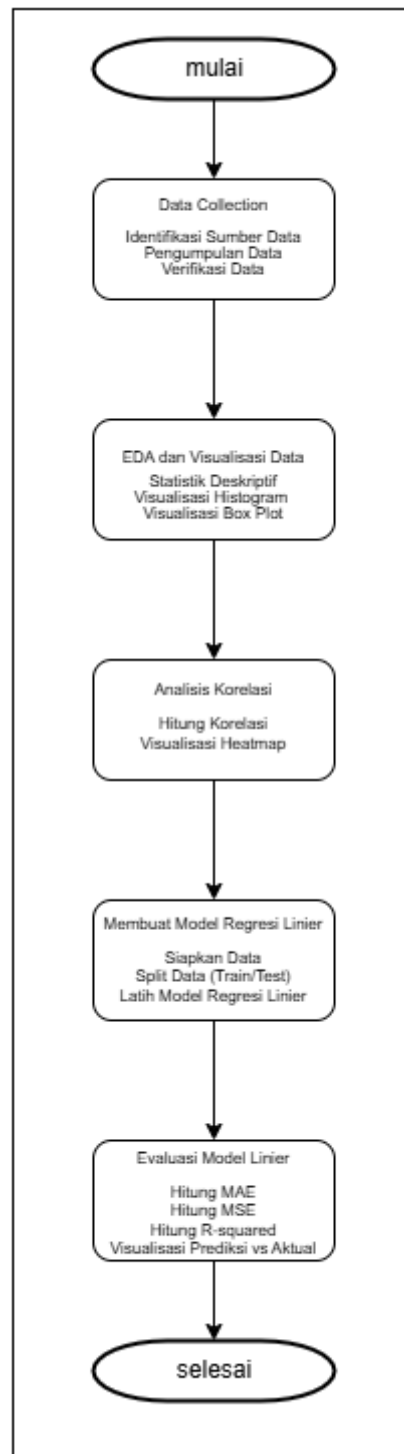
#### 4. Membuat Model Regresi Linier

- Tujuan: Memprediksi jumlah wisatawan di masa depan berdasarkan data historis.
- Proses:
  - Model Regresi Linier Sederhana: Membuat model regresi linier sederhana untuk memprediksi jumlah wisatawan berdasarkan satu variabel independen (misalnya, bulan).
  - Model Regresi Linier Berganda: Membuat model regresi linier berganda dengan beberapa variabel independen (misalnya, bulan dan negara asal).
  - Training dan Testing: Membagi data menjadi set pelatihan dan set pengujian, melatih model pada data pelatihan, dan menguji performa model pada data pengujian.
  - Insight: Mengidentifikasi variabel yang memiliki pengaruh signifikan terhadap jumlah wisatawan dan membuat prediksi berdasarkan model.

#### 5. Evaluasi Model Linier

- Tujuan: Memprediksi jumlah wisatawan di masa depan berdasarkan data historis.
- Proses:
  - Model Regresi Linier Sederhana: Membuat model regresi linier sederhana untuk memprediksi jumlah wisatawan berdasarkan satu variabel independen (misalnya, bulan).
  - Model Regresi Linier Berganda: Membuat model regresi linier berganda dengan beberapa variabel independen (misalnya, bulan dan negara asal).
  - Training dan Testing: Membagi data menjadi set pelatihan dan set pengujian, melatih model pada data pelatihan, dan menguji performa model pada data pengujian.
  - Insight: Mengidentifikasi variabel yang memiliki pengaruh signifikan terhadap jumlah wisatawan dan membuat prediksi berdasarkan model.

- Flowchart



## 2.2. Dataset

Data yang digunakan diperoleh dari situs web resmi Badan Pusat Statistik Indonesia. Data yang diambil dan dipergunakan untuk eksperimen yakni data terbaru 2024. <https://www.bps.go.id/id/statistics-table/2/MTQ3MCMY/jumlah-kunjungan-wisatawan-mancanegara-per-bulan-menurut-kebangsaan--kunjungan-.html> .

Kemudian dataset jadi yang siap dipakai:

<https://drive.google.com/file/d/1rOxMIFtEfFMpsG17pTbbYy0gHqOxMBjH/view?usp=sharing> .

Dataset yang upload berisi data jumlah wisatawan berdasarkan kebangsaan dari Januari hingga Desember, serta jumlah tahunan. Berikut adalah beberapa detail mengenai dataset ini:

- Jumlah Baris dan Kolom: Dataset ini memiliki 249 baris dan 14 kolom.
- Kolom:
  - Kebangsaan: Kebangsaan wisatawan.
  - Januari, Februari, Maret, April, Mei: Jumlah wisatawan per bulan dari Januari hingga Mei (tipe data int64).
  - Juni, Juli, Agustus, September, Oktober, November, Desember, Tahunan: Data ini berisi simbol - dan tipe data object.
- Contoh Data:
  - Kebangsaan: Brunei Darussalam, Malaysia, Philippines, Singapore, Thailand
  - Jumlah wisatawan per bulan (Januari hingga Mei) tercatat dengan nilai numerik.
  - Kolom dari Juni hingga Desember, serta kolom Tahunan berisi simbol -.

```
# Membaca data dari file CSV
df = pd.read_csv('dataset_wisatawan_2024.csv')

# Menampilkan beberapa baris pertama data
print("Head of the dataset:")
print(df.head())
```

Head of the dataset:

	Kebangsaan	Januari	Februari	Maret	April	Mei	Juni	Juli	\
0	Brunei Darussalam	747	1251	999	857	1359	-	-	
1	Malaysia	155213	218057	160269	170644	200070	-	-	
2	Philippines	16937	18367	20469	17726	19462	-	-	
3	Singapore	87248	114301	120040	81225	111021	-	-	
4	Thailand	8449	8666	8691	9791	11081	-	-	

	Agustus	September	Oktober	November	Desember	Tahunan
0	-	-	-	-	-	-
1	-	-	-	-	-	-
2	-	-	-	-	-	-
3	-	-	-	-	-	-
4	-	-	-	-	-	-



## 2.3. EDA

1. Memahami Data
  - Mengimpor Dataset: Membaca data dari file CSV.
  - Memeriksa Struktur Data: Menggunakan `info()` untuk melihat tipe data dan jumlah nilai non-null di setiap kolom.
  - Melihat Sampel Data: Menggunakan `head()` untuk melihat beberapa baris pertama dari dataset.
2. Data Cleaning (Pembersihan Data)
  - Mengidentifikasi Missing Values: Mengecek nilai yang hilang atau tidak lengkap dalam dataset. (dengan fungsi `isnull().sum()`)
  - Menangani Missing Values: Mengisi nilai yang hilang atau mengubah nilai simbol - menjadi NaN, kemudian memutuskan apakah akan mengisi atau menghapusnya. (dengan fungsi `replace()`)
  - Memperbaiki Tipe Data: Mengubah tipe data yang tidak sesuai (misalnya, konversi kolom yang seharusnya numerik tapi bertipe object). (dengan fungsi `pd.to.numeric()`)
  - Menghilangkan Duplikasi: Memeriksa dan menghapus baris duplikat dalam data. (dengan fungsi `drop_duplicates()`)
  - Menghilangkan beberapa baris data yang tidak diperlukan. (dengan fungsi `df.index` untuk menyeleksi baris berapa, dan `drop(index)` untuk menghapus baris berdasarkan index berapa)
3. Data Transformation (Transformasi Data)
  - Normalisasi dan Standarisasi: Mengubah skala data jika diperlukan untuk analisis lebih lanjut.
  - Membuat Fitur Baru: Menambah kolom baru berdasarkan kolom yang ada (misalnya, membuat kolom total tahunan dari jumlah bulanan). (dengan fungsi `iloc[]`)
4. Descriptive Statistics (Statistik Deskriptif)
  - Statistik Dasar: Menghitung mean, median, mode, standar deviasi, dan statistik deskriptif lainnya untuk setiap kolom. (dengan fungsi `describe()`)
  - Distribusi Data: Menganalisis distribusi data untuk setiap kolom (misalnya menggunakan histogram atau boxplot).
5. Data Visualization (Visualisasi Data)
  - Plot Univariate: Membuat visualisasi untuk satu variabel, seperti histogram, bar plot, dan boxplot. (`hist()`)
  - Plot Bivariate: Membuat visualisasi untuk dua variabel, seperti scatter plot, line plot, dan bar plot. (`plot()`)
  - Plot Multivariate: Membuat visualisasi untuk lebih dari dua variabel, seperti pair plot atau heatmap korelasi. (`sns.heatmap()`)

6. Data Analysis (Analisis Data)

- Mencari Pola dan Tren: Mengidentifikasi pola atau tren dalam data (misalnya, tren musiman atau pola tertentu dalam data waktu).
- Analisis Korelasi: Menghitung dan menganalisis korelasi antara variabel untuk memahami hubungan antar variabel. (`corr()`)
- Segmentasi Data: Mengelompokkan data berdasarkan kategori tertentu untuk analisis lebih lanjut.

### 3. Eksperimen

- Proses Eksperimen

1. Data Collection

Pengumpulan data dari file CSV yang telah diunggah. Data berisi jumlah wisatawan per negara dari Januari hingga Desember serta jumlah tahunan.

```
# Membaca data dari file CSV
df = pd.read_csv('dataset_wisatawan_2024.csv')

# Menampilkan beberapa baris pertama data
print("Head of the dataset:")
print(df.head())
```

2. Exploratory Data Analysis (EDA)

Memahami struktur, karakteristik, dan pola dalam data dengan menggunakan statistik deskriptif dan visualisasi.

- Memahami struktur data

```
# Informasi dasar tentang dataset
print("\nInfo of the dataset:")
print(df.info())
print("\n-----\n")

# Melihat isi dari DataFrame
print("\nContent of the dataset:")
print(df)
print("\n-----\n")

# Statistik deskriptif
print("\nDescriptive statistics of the dataset:")
print(df.describe())
print("\n-----\n")

# Melihat tipe data dari setiap kolom
print("\nData types of the columns:")
df.dtypes
```

## - Data Cleaning

```
UAS BDPA 2024_Kelompok 6.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
RAM 100% Disk 100% + Gemini

[ ] # Menyeleksi data yang dibutuhkan saja

# Menghapus indeks yang tidak dipakai
# indeks dihapus = df.index[df.index > 9]
indeks_dihapus = df.index[df.index > 38]

# Menghapus baris berdasarkan dari indeks yang dipilih
baris_dihapus = df.drop(index=indeks_dihapus)

# Menghapus kolom yang tidak diperlukan
# Menghapus juga kolom yang nilainya null atau 0 atau - (strip)
# kolom dihapus = baris_dihapus.drop(columns=['Juni', 'Juli', 'Agustus', 'September', 'Oktober', 'November', 'Desember', 'Tahunan'])
# data dipakai = kolom_dihapus
# print("Data dipakai:\n")
# print(data_dipakai)

data_dipakai = baris_dihapus
print("Data dipakai:\n")
print(data_dipakai)
print("\n-----\n")

# hapus total asen
hapus_asen = data_dipakai.index[data_dipakai.index == 10]
hapus_baris = data_dipakai.drop(index=hapus_asen)
data_final = hapus_baris
print("Data tidak ada baris asen:\n")
print(data_final)
print("\n-----\n")

# untuk blank plot
hapus_nilai_strip = data_final.drop(columns=['Juni', 'Juli', 'Agustus', 'September', 'Oktober', 'November', 'Desember', 'Tahunan'])
data_clean = hapus_nilai_strip
print("Data tidak ada nilai strip:\n")
print(data_clean)
print("\n-----\n")

print("Data tidak ada kolom kebangsaan:\n")
# digunakan untuk visualisasi heatmap
hapus_index_kolom_kebangsaan = data_final.drop(columns=['Kebangsaan'])
print(hapus_index_kolom_kebangsaan)
```

## - Descriptive Statistics

```
# Statistik deskriptif
print("\nDescriptive statistics of the dataset:")
print(df.describe())
print("\n-----\n")
```

## - Visualisasi Data

```
[ ] # Visualisasi jumlah wisatawan bulanan dari beberapa negara
plt.figure(figsize=(15, 8))

for country in data_final['Kebangsaan'].unique()[:5]:
    subset = data_final[data_final['Kebangsaan'] == country]
    plt.plot(subset.columns[1:-8], subset.iloc[0, 1:-8], label=country)

plt.title('Jumlah Wisatawan Bulanan dari Beberapa Negara')
plt.xlabel('Bulan')
plt.ylabel('Jumlah Wisatawan')
plt.legend()
plt.xticks(rotation=45)
plt.show()
```

```
[ ] # Visualisasi 1: Distribusi wisatawan per bulan
plt.figure(figsize=(14, 8))

df_sum = data_final.iloc[:, 1:-8].sum()
df_sum.plot(kind='bar')
plt.title('Distribusi Wisatawan per Bulan')
plt.xlabel('Bulan')
plt.ylabel('Jumlah Wisatawan')
plt.xticks(rotation=45)
plt.show()
```

### 3. Analisis Korelasi

Mengidentifikasi hubungan antar variabel dalam dataset untuk menentukan variabel yang mungkin saling berpengaruh.

```
▶ # 3. Analisis Korelasi

# Menghitung matriks korelasi
correlation_matrix = data_final.iloc[:, 1:-8].corr()

# Menampilkan matriks korelasi
print(correlation_matrix)
```

```
[ ] # Visualisasi 3: Heatmap Korelasi
# Analisis Korelasi
correlation_matrix = data_final.iloc[:, 1:-8].corr()

# Visualisasi matriks korelasi
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Matriks Korelasi antara Jumlah Wisatawan Bulanan dan Tahunan')
plt.show()
```

#### 4. Membuat Model Regresi Linier

Membangun model regresi linier untuk memprediksi jumlah wisatawan berdasarkan data historis.

```
[ ] # 4. Membuat Model Regresi Linier Sederhana

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Membuat model regresi linier sederhana
X_sederhana = data_final[['Februari']]
y = data_final['Januari']

# Membagi data menjadi training dan testing set
X_train_sederhana, X_test_sederhana, y_train_sederhana, y_test_sederhana = train_t

# Melatih model
model_sederhana = LinearRegression()
model_sederhana.fit(X_train_sederhana, y_train_sederhana)

# Memprediksi nilai
y_pred_sederhana = model_sederhana.predict(X_test_sederhana)

# Evaluasi model
mse = mean_squared_error(y_test_sederhana, y_pred_sederhana)
r2 = r2_score(y_test_sederhana, y_pred_sederhana)

print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
[ ] # 5. Membuat Model Regresi Berganda

# Membuat model regresi linier berganda
X_berganda = data_final[['Februari', 'Maret', 'April', 'Mei']]
y = data_final['Januari']

# Membagi data menjadi training dan testing set
X_train_berganda, X_test_berganda, y_train_berganda, y_test_berganda = train_test_

# Melatih model
model_berganda = LinearRegression()
model_berganda.fit(X_train_berganda, y_train_berganda)

# Memprediksi nilai
y_pred_berganda = model_berganda.predict(X_test_berganda)

# Evaluasi model
mse = mean_squared_error(y_test_berganda, y_pred_berganda)
r2 = r2_score(y_test_berganda, y_pred_berganda)

print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

## 5. Evaluasi Model Linier

Mengevaluasi kinerja model dengan metrik seperti Mean Absolute Error (MAE), Mean Squared Error (MSE), dan R-squared ( $R^2$ ).

```
[ ] # 5. Evaluasi Model Linier

# Visualisasi Prediksi vs Nilai Aktual untuk regresi linier sederhana
plt.figure(figsize=(10, 6))
plt.scatter(y_test_sederhana, y_pred_sederhana, edgecolors=(0, 0, 0))
plt.plot([y_test_sederhana.min(), y_test_sederhana.max()], [y_test_sederhana.min(), y_test_sederhana.max()], 'k--', lw=2)
plt.xlabel('Nilai Aktual')
plt.ylabel('Nilai Prediksi')
plt.title('Prediksi vs Nilai Aktual (Regresi Linier Sederhana)')
plt.show()
print("\n-----\n")

# Urutkan data berdasarkan nilai aktual untuk visualisasi yang lebih jelas
sorted_indices_sederhana = np.argsort(y_test_sederhana)
y_test_sederhana_sorted = y_test_sederhana.iloc[sorted_indices_sederhana].values
y_pred_sederhana_sorted = y_pred_sederhana[sorted_indices_sederhana]

# Visualisasi Prediksi vs Nilai Aktual (Line Chart)
plt.figure(figsize=(14, 6))
plt.plot(y_test_sederhana_sorted, label='Nilai Aktual', color='blue', linestyle='-', marker='o')
plt.plot(y_pred_sederhana_sorted, label='Nilai Prediksi', color='red', linestyle='--', marker='x')
plt.xlabel('Indeks')
plt.ylabel('Nilai')
plt.title('Prediksi vs Nilai Aktual (Regresi Linier Sederhana) - Line Chart')
plt.legend()
plt.show()
```

```

# Visualisasi Prediksi vs Nilai Aktual untuk regresi linier berganda
plt.figure(figsize=(10, 6))
plt.scatter(y_test_berganda, y_pred_berganda, edgecolors=(0, 0, 0))
plt.plot([y_test_berganda.min(), y_test_berganda.max()], [y_test_berganda.min(), y_test_berganda.max()], 'k--', lw=2)
plt.xlabel('Nilai Aktual')
plt.ylabel('Nilai Prediksi')
plt.title('Prediksi vs Nilai Aktual (Regresi Linier Berganda)')
plt.show()
print("\n----\n")

# Urutkan data berdasarkan nilai aktual untuk visualisasi yang lebih jelas
sorted_indices_berganda = np.argsort(y_test_berganda)
y_test_berganda_sorted = y_test_berganda.iloc[sorted_indices_berganda].values
y_pred_berganda_sorted = y_pred_berganda[sorted_indices_berganda]

# Visualisasi Prediksi vs Nilai Aktual (Line Chart)
plt.figure(figsize=(14, 6))
plt.plot(y_test_berganda_sorted, label='Nilai Aktual', color='blue', linestyle='-', marker='o')
plt.plot(y_pred_berganda_sorted, label='Nilai Prediksi', color='red', linestyle='--', marker='x')
plt.xlabel('Indeks')
plt.ylabel('Nilai')
plt.title('Prediksi vs Nilai Aktual (Regresi Linier Berganda) - Line Chart')
plt.legend()
plt.show()

```

## 6. Rangkuman

- Data Collection: Mengimpor data dari file CSV.
- EDA dan Visualisasi Data: Memahami struktur data, membersihkan data, dan melakukan visualisasi.
- Analisis Korelasi: Mengidentifikasi hubungan antar variabel.
- Membuat Model Regresi Linier: Membangun model untuk memprediksi jumlah wisatawan.
- Evaluasi Model Linier: Mengevaluasi kinerja model menggunakan matrik evaluasi.



- Library dan Tools

## Library

### 1. Pandas

Pandas adalah library Python yang sangat populer untuk manipulasi dan analisis data. Pandas menyediakan struktur data yang fleksibel dan efisien untuk menangani data berlabel, seperti data frame.

Penggunaan dalam Eksperimen:

- Membaca dataset dari file CSV.
- Memeriksa struktur data dan informasi dasar.
- Mengganti nilai yang hilang atau tidak valid.
- Menghitung statistik deskriptif.
- Mengubah dan memanipulasi data.

### 2. Matplotlib

Matplotlib adalah library untuk membuat visualisasi data di Python. Library ini sangat berguna untuk membuat plot, grafik, dan diagram dengan berbagai jenis dan gaya.

Penggunaan dalam Eksperimen:

- Membuat histogram untuk distribusi data.
- Membuat line plot untuk melihat tren data.

### 3. Seaborn

Seaborn adalah library visualisasi data yang dibangun di atas Matplotlib. Seaborn menyediakan antarmuka tingkat tinggi yang lebih mudah digunakan untuk membuat visualisasi statistik yang menarik dan informatif.

Penggunaan dalam Eksperimen:

Membuat heatmap untuk menunjukkan korelasi antar variabel.

### 4. Scikit-learn

Scikit-learn adalah library machine learning di Python yang menyediakan berbagai alat untuk pemodelan dan analisis data, termasuk algoritma klasifikasi, regresi, klustering, dan matrik evaluasi.

Penggunaan dalam Eksperimen:

- Membagi dataset menjadi data pelatihan dan pengujian.
- Membangun model regresi linier.
- Mengevaluasi model menggunakan metrik evaluasi.

## 5. NumPy

NumPy adalah library dasar untuk komputasi ilmiah di Python. NumPy menyediakan array n-dimensi, serta berbagai fungsi matematika dan aljabar linier.

Penggunaan dalam Eksperimen:

Sering digunakan secara internal oleh Pandas dan Scikit-learn untuk operasi matematika dan komputasi.

Tools

Google Colab, atau Google Colaboratory, adalah lingkungan pemrograman interaktif berbasis cloud yang memungkinkan untuk menulis dan menjalankan kode Python langsung dari browser.

## 4. Hasil dan Evaluasi

- Hasil
  - Data Collection: Dataset berhasil diimpor dan diperiksa.
  - Exploratory Data Analysis (EDA) dan Visualisasi Data: Struktur data dipahami, data dibersihkan, dan berbagai visualisasi menunjukkan distribusi data dan hubungan antar variabel.
  - Analisis Korelasi: Korelasi antara variabel bulanan dapat diidentifikasi.
  - Membuat Model Regresi Linier: Model dibangun untuk memprediksi bulan januari berdasarkan data bulanan.
  - Evaluasi Model Linier: Kinerja model dievaluasi, dengan hasil menunjukkan kemampuan model untuk memprediksi bulan januari.

- Evaluasi

Hasil yang didapat:

1. Regresi Linier Sederhana
  - Mean Squared Error (MSE): 86,379,861.88
  - R-squared ( $R^2$ ): 0.7899
2. Regresi Linier Berganda
  - Mean Squared Error (MSE): 34,978,317.59
  - R-squared ( $R^2$ ): 0.9149

Evaluasi	Regresi Linier Sederhana	Regresi Linier Berganda
Akurasi Model	Model ini memiliki MSE yang cukup tinggi dan $R^2$ sekitar 0.79. Ini menunjukkan bahwa model sederhana ini mampu menjelaskan sekitar 79% variabilitas dalam data, tetapi masih ada ruang untuk perbaikan karena MSE yang relatif tinggi.	Model ini memiliki MSE yang lebih rendah dan $R^2$ sekitar 0.91. Ini menunjukkan bahwa model regresi berganda lebih akurat dalam menjelaskan variabilitas data (sekitar 91%) dan menghasilkan prediksi yang lebih dekat dengan nilai aktual dibandingkan dengan model sederhana.
Efektivitas Model	Menggunakan hanya satu variabel independen (Februari) untuk memprediksi jumlah wisatawan tahunan memberikan hasil yang cukup baik, namun tidak optimal. Model ini memberikan gambaran umum, tetapi tidak menangkap semua variabilitas dalam data.	Dengan menggunakan beberapa variabel independen (Februari, Maret, April, Mei), model ini dapat menangkap lebih banyak informasi dan variabilitas dalam data. Hal ini tercermin dari MSE yang lebih rendah dan $R^2$ yang lebih tinggi, yang menunjukkan bahwa model ini lebih efektif dalam memprediksi jumlah wisatawan.
Penggunaan Variabel	Mengandalkan satu variabel mungkin tidak cukup untuk menangkap kompleksitas dalam dataset wisatawan.	Menggunakan beberapa variabel memberikan gambaran yang lebih lengkap dan meningkatkan kemampuan prediksi model.

### 3. Kesimpulan dan Kontribusi

#### a. Kesimpulan

Regresi Linier Berganda direkomendasikan karena cenderung memberikan prediksi yang lebih akurat dibanding metode lain, dengan nilai Mean Squared Error (MSE) yang rendah dan koefisien determinasi ( $R^2$ ) yang tinggi. Analisis lebih lanjut terhadap variabel independen yang digunakan sangat penting untuk memastikan relevansi dan kontribusi positif terhadap model.

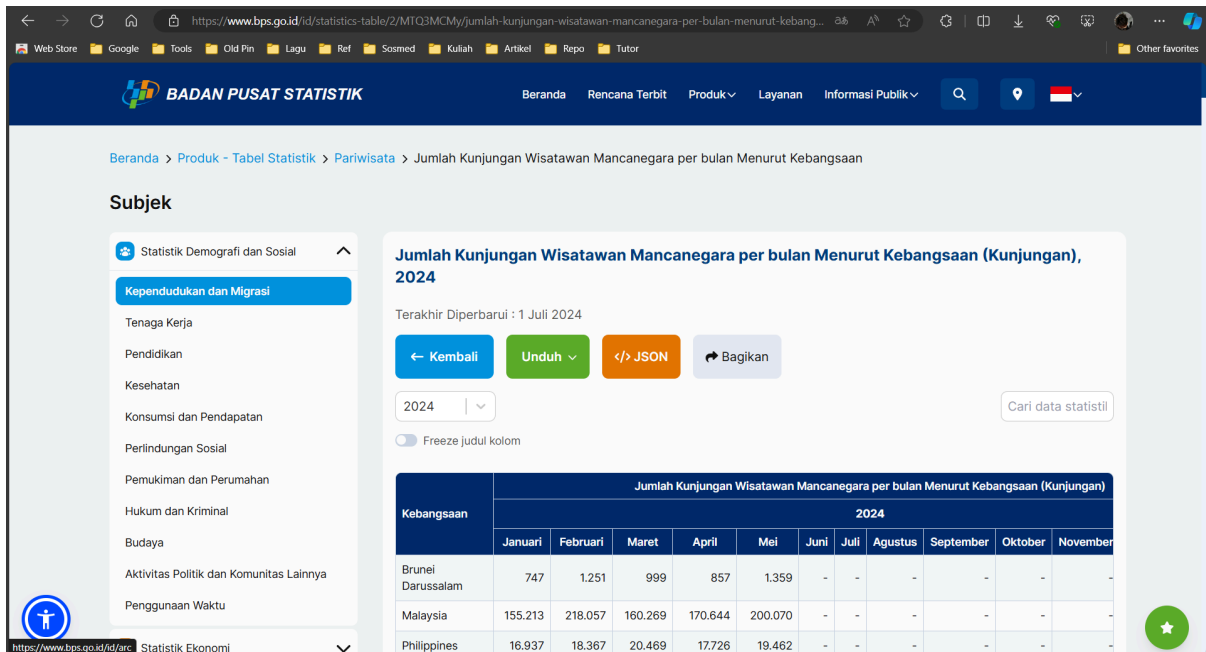
Jika diperlukan, pertimbangkan penambahan variabel baru yang dapat meningkatkan kinerja prediktif. Validasi model melalui teknik cross-validation sangat dianjurkan guna memastikan bahwa model tidak overfitting dan dapat diandalkan dalam memprediksi data baru. Selain itu, pastikan kebersihan data dengan mengeliminasi nilai yang hilang atau outlier yang signifikan yang mungkin mempengaruhi hasil model.

#### b. Kontribusi

- Bakri Ahmad Ridhwan: Melakukan koding eksperimen pada dataset menggunakan google colaboratory, dan membuat dokumen laporan.
- Choirul Affan Adi Putra: Melakukan koding eksperimen pada dataset menggunakan google colaboratory, dan membuat dashboard dari eksperimen menggunakan flask.
- Dhimas Adhiyaksa Davariza Yudhaswara: Melakukan koding eksperimen pada dataset menggunakan google colaboratory, dan membuat poster dari eksperimen.

## 4. Lampiran

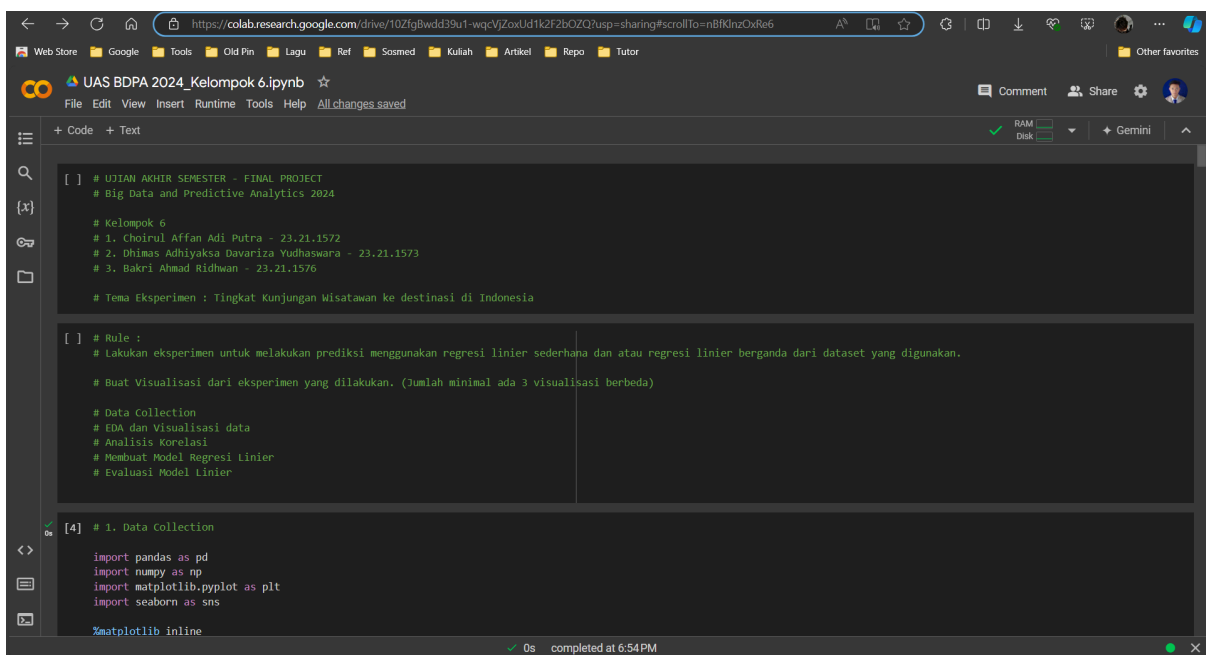
- Dataset: [Jumlah Kunjungan Wisatawan Mancanegara per bulan Menurut Kebangsaan - Tabel Statistik - Badan Pusat Statistik Indonesia \(bps.go.id\)](https://www.bps.go.id/statistics-table/2/MIQ3MCMY/jumlah-kunjungan-wisatawan-mancanegara-per-bulan-menurut-kebangsaan)



The screenshot shows the BPS website with the following table data:

Kebangsaan	Jumlah Kunjungan Wisatawan Mancanegara per bulan Menurut Kebangsaan (Kunjungan)										
	2024										
	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus	September	Oktober	November
Brunel Darussalam	747	1.251	999	857	1.359	-	-	-	-	-	-
Malaysia	155.213	218.057	160.269	170.644	200.070	-	-	-	-	-	-
Philippines	16.937	18.367	20.469	17.726	19.462	-	-	-	-	-	-

- Notebook / link google collab: <https://colab.research.google.com/drive/10ZfgBwdd39u1-wqcVjZoxUd1k2F2bOZQ?usp=sharing>



The screenshot shows a Google Colab notebook with the following content:

```
# UJIAN AKHIR SEMESTER - FINAL PROJECT
# Big Data and Predictive Analytics 2024

# Kelompok 6
# 1. Choirul Affan Adi Putra - 23.21.1572
# 2. Dhinias Adhiyaksa Davariza Yudhaswara - 23.21.1573
# 3. Bakri Ahmad Ridhwan - 23.21.1576

# Tema Eksperimen : Tingkat Kunjungan Wisatawan ke destinasi di Indonesia

# Rule :
# Lakukan eksperimen untuk melakukan prediksi menggunakan regresi linier sederhana dan atau regresi linier berganda dari dataset yang digunakan.
# Buat Visualisasi dari eksperimen yang dilakukan. (jumlah minimal ada 3 visualisasi berbeda)

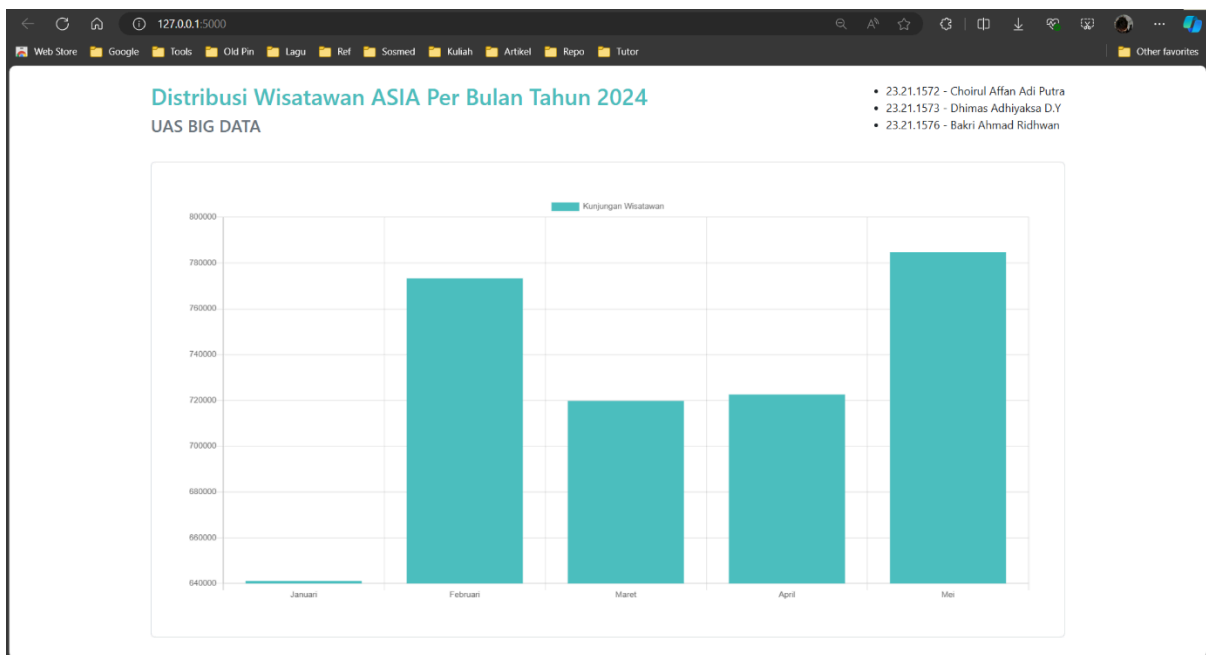
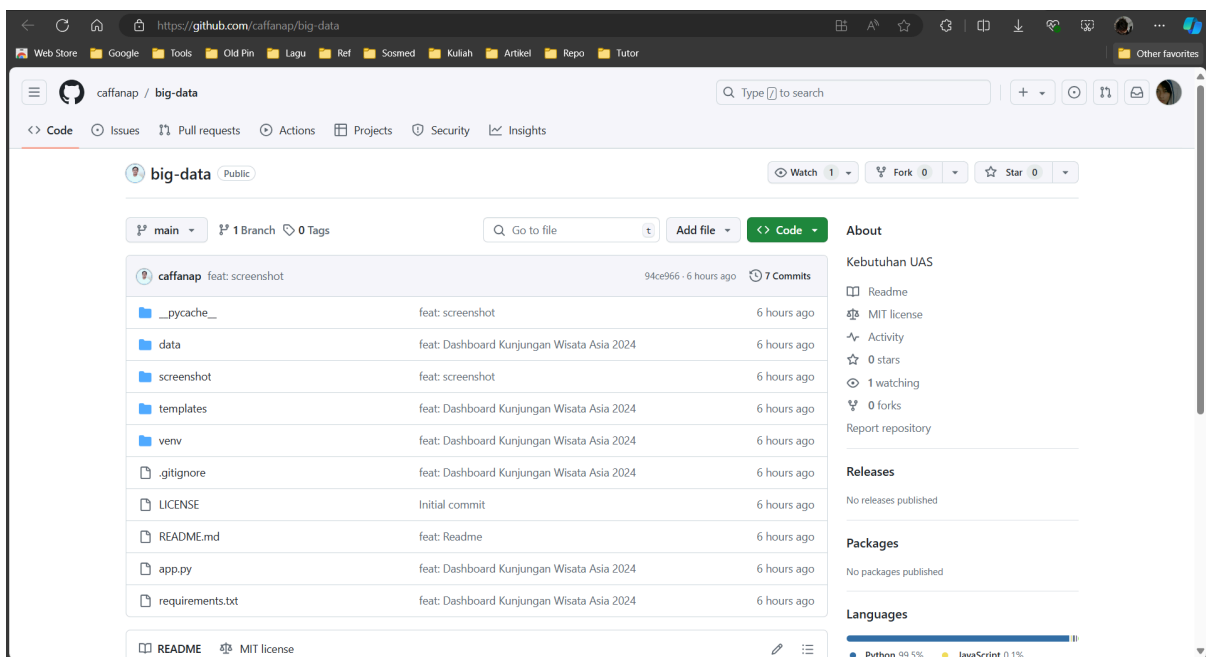
# Data Collection
# EDA dan Visualisasi data
# Analisis Korelasi
# Membuat Model Regresi Linier
# Evaluasi Model Linier

[4] # 1. Data Collection

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
```

- Dashboard: [caffanap/big-data: Kebutuhan UAS \(github.com\)](https://github.com/caffanap/big-data)

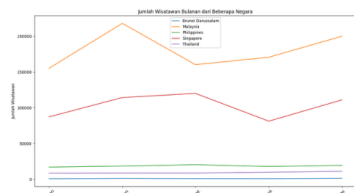


- Poster: [Production Performance Data Infographic - Infographic \(canva.com\)](https://www.canva.com)

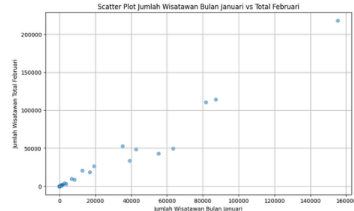
### Tingkat Kunjungan Wisatawan ke Destinasi di Indonesia

Pariwisata merupakan sektor ekonomi yang memiliki peran penting bagi banyak negara, termasuk Indonesia. Sebagai negara kepulauan, Indonesia memiliki kekayaan alam, budaya, dan sejarah yang menarik bagi wisatawan mancanegara. Oleh karena itu, memahami tren kunjungan wisatawan dan faktor-faktor yang mempengaruhinya menjadi krusial dalam pengembangan sektor pariwisata.

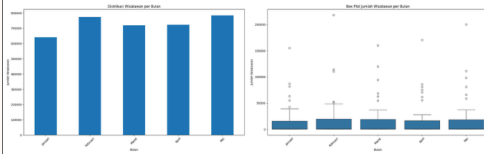
#### Line Chart



#### Scatter Plot



#### Bar Chart & Box Plot



Berdasarkan Dataset "Jumlah Kunjungan Wisatawan Mancanegara per bulan Menurut Kebangsaan (Kunjungan), 2023" dari Badan Pusat Statistik berisi informasi tentang kunjungan wisatawan mancanegara ke berbagai destinasi di Indonesia. Data ini mencakup periode waktu tertentu dan mencatat jumlah kunjungan, profil wisatawan, serta faktor-faktor lain yang relevan. Dengan menganalisis dataset ini, kita dapat memperoleh wawasan yang berharga untuk pengambilan keputusan di sektor pariwisata.

👤 Bakri Ahmad Ridhwan 23.21.1576

👤 Choirul Affan A.P 23.21.1572

👤 Dhimas A.D.Y 23.21.1573

### Tingkat Kunjungan Wisatawan ke Destinasi di Indonesia

#### Kesimpulan

##### Regresi Linier Sederhana

- Mean Squared Error (MSE): 86,379,881.88
- R-squared ( $R^2$ ): 0.7899

##### Akurasi Model:

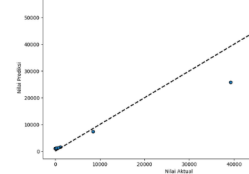
Regresi Linier Sederhana: Model ini memiliki MSE yang cukup tinggi dan  $R^2$  sekitar 0.79. Ini menunjukkan bahwa model sederhana ini mampu menjelaskan sekitar 79% variabilitas dalam data, tetapi masih ada ruang untuk perbaikan karena MSE yang relatif tinggi.

##### Efektivitas Model:

Regresi Linier Sederhana: Menggunakan hanya satu variabel independen (Februari) untuk memprediksi jumlah wisatawan tahunan memberikan hasil yang cukup baik, namun tidak optimal. Model ini memberikan gambaran umum, tetapi tidak menangkap semua variabilitas dalam data.

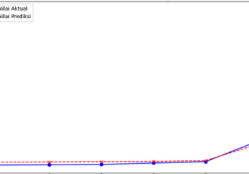
#### Prediksi vs Nilai Aktual (Regresi Linear Sederhana)

Prediksi vs Nilai Aktual (Regresi Linear Sederhana)



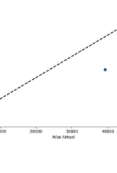
#### Prediksi vs Nilai Aktual (Regresi Linear Sederhana) - Line Chart

Prediksi vs Nilai Aktual (Regresi Linear Sederhana) - Line Chart



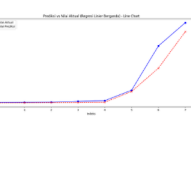
#### Prediksi vs Nilai Aktual (Regresi Linear Berganda)

Prediksi vs Nilai Aktual (Regresi Linear Berganda)



#### Prediksi vs Nilai Aktual (Regresi Linear Berganda) Line Chart

Prediksi vs Nilai Aktual (Regresi Linear Berganda) Line Chart



##### Regresi Linier Berganda

- Mean Squared Error (MSE): 34,978,317.59
- R-squared ( $R^2$ ): 0.9149

##### Akurasi Model:

Regresi Linier Berganda: Model ini memiliki MSE yang lebih rendah dan  $R^2$  sekitar 0.91. Ini menunjukkan bahwa model regresi berganda lebih akurat dalam menjelaskan variabilitas data (sekitar 91%) dan menghasilkan prediksi yang lebih dekat dengan nilai aktual dibandingkan dengan model sederhana.

##### Efektivitas Model:

Regresi Linier Berganda: Dengan menggunakan beberapa variabel independen (Februari, Maret, April, Mei), model ini dapat menangkap lebih banyak informasi dan variabilitas dalam data. Hal ini tercermin dari MSE yang lebih rendah dan  $R^2$  yang lebih tinggi, yang menunjukkan bahwa model ini lebih efektif dalam memprediksi jumlah wisatawan.

##### Penggunaan Variabel:

Regresi Linier Sederhana: Mengandalkan satu variabel mungkin tidak cukup untuk menangkap kompleksitas dalam dataset wisatawan.

Regresi Linier Berganda: Menggunakan beberapa variabel memberikan gambaran yang lebih lengkap dan meningkatkan kemampuan prediksi model.

👤 Bakri Ahmad Ridhwan 23.21.1576

👤 Choirul Affan A.P 23.21.1572

👤 Dhimas A.D.Y 23.21.1573