Penerapan Data Science pada Dataset Olympics

Gregorius Kennard Taruna[#1], Setia Budi[*2]

Program Studi Sistem Informasi, Universitas Kristen Maranatha
Jl. Prof Drg Surya Sumantri 65, Kota Bandung, Indonesia
[1]1873012@maranatha.ac.id
[2] setia.budi@it.maranatha.edu

Abstract — **The Olympics is an international sporting event with many sports participated by Indonesia, Japan, America, and others. One example of this year's Olympics is the 2021 Tokyo Olympics. In the Tokyo 2021 Olympics dataset, several data will be used. This dataset will be processed using the Python Pandas programming language. This programming language will be created using an open-source web application called Jupyter Notebook. In a separate dataset, there is information that can be viewed, namely Athlete data, Coach data, EntriesGender data, Medal data, and Team data. In other datasets, there are Hosts data, Athletes data, Medals data, and Results data. The source of both datasets comes from the Kaggle website. From this dataset, it can be analyzed, here are some examples of analysis, namely looking for unique athletes. In addition, look for the GOAT in the sport that is always there. From this analysis, it produces data that becomes information on the results of exploration and visualization.**
Keywords— **Olympics, dataset, python pandas, data**

I. PENDAHULUAN

Olimpiade adalah salah satu pertandingan olahraga internasional yang diadakan dari tahun 1896 sampai tahun 2021. Ajang olahraga ini biasa diadakan dalam empat tahun sekali. Salah satu contoh Olimpiade pada tahun ini yaitu Olimpiade Tokyo 2021. Pada awalnya, Olimpiade ini akan digelar pada Juli 2020 tetapi karena Virus COVID-19 maka diadakan pada Juli 2021. Berikut beberapa negara yang mengikuti ajang olahraga ini adalah Indonesia, Italia, Belanda, Jepang, dan lain-lain. Dalam Olimpiade Tokyo 2021 Indonesia mengikuti beberapa cabang olahraga antara lain yaitu bulu tangkis, angkat besi, atletik, dayung, renang, dan lain-lain.

Dalam dataset Olimpiade Tokyo 2021 dan Olimpiade sebelumnya terdapat data yang akan digunakan. Berikut terdapat 5 dataset pada Olimpiade Tokyo 2021 dan 4 dataset pada Olimpiade tahun 1896 – 2021. Dalam dataset Olimpiade Tokyo 2021 terdapat data athlete, coaches, entriesgender, medals, dan teams. Sedangkan pada dataset Olimpiade tahun 1896 – 2021 terdapat data athletes, hosts, medals, dan results.

Dari dataset tersebut dapat kita eksplorasi lebih dalam lagi. Contoh dari eksplorasi yang bisa dilakukan yaitu mencari jumlah atlet dari setiap negara. Berikutnya kita dapat mencari negara dengan jumlah medali. Selain itu ada perbandingan jumlah gender dalam partisipasi di setiap cabang olahraga. Sebenarnya, masih banyak yang dapat dieksplorasi dari dataset tersebut dan akan dibahas pada sub-bab selanjutnya.

II. KAJIAN TEORI

A. Olimpiade

Olimpiade merupakan salah satu pertandingan olahraga internasional yang memiliki banyak cabang olahraga yang diikuti ribuan atlet. Pertama kali Olimpiade dilaksanakan di Yunani pada tiga ribu tahun lalu. Dahulu Olimpiade ini dilaksanakan untuk menghormati dewa Zeus yang disembah oleh orang Yunani Kuno. Sejak dahulu Olimpiade ini diadakan setiap empat tahun sekali selama musim panas [1].

Olimpiade Tokyo 2021 merupakan ajang olahraga yang diikuti dari berbagai negara dan diselenggarakan di Jepang. Olimpiade ini bermula diselenggarakan pada tahun 2020 tetapi ditunda karena terdapat wabah COVID-19 dan diadakan kembali pada tahun 2021. Jepang telah melakukan dua kali dalam Olimpiade yaitu pada tahun 1940 dan tahun 2020 [2]. Dalam Olimpiade ini terdapat 204 negara yang berpartisipasi dan terdapat sekitar 20 juta pengunjung. Dalam pertandingan ini akan diikuti 11.090 atlet dan diperkirakan akan disediakan makanan sebesar 14 juta makanan [3].

B. Data

Data adalah sekumpulan informasi atau fakta yang dikumpulkan peneliti melalui sumber-sumber tertentu [4]. Dalam hal penelitian data dapat membantu menjawab pertanyaan peneliti. Lalu jenis data dapat dibagi menjadi 2 yaitu data primer dan

sekunder. Pengertian dari data primer adalah sekumpulan data yang berasal dari sumber asli. Sedangkan data sekunder berasal dari informasi yang sudah diolah oleh pihak lain [5]. Data sekunder yang akan digunakan dalam eksplorasi kali ini menggunakan metode Exploratory Data Analysis.

C. Exploratory Data Analysis

EDA atau *Exploratory Data Analysis* adalah metode eksplorasi yang digunakan untuk menganalisis kumpulan data yang biasanya menggunakan metode visual. EDA berguna untuk memahami data dan memahami variabel. Metode ini sangat baik untuk menemukan data yang berguna dalam menyeleksi model statistik yang tepat [6].

D. Jupyter Notebook

Jupyter Notebook merupakan *software* berupa aplikasi *web open-source* yang penggunanya dapat menggabungkan *live code, markdown*, gambar, plot, dan lainnya dalam satu dokumen. Notebook ini dibuat oleh Jupyter dan evolusi dari IPython. JuPyteR bermula dari dukungan bahasa pemrograman Julia, Python dan R. Selain itu sekarang telah didukung bahasa pemrograman lainnya seperti Scala, Haskell, dan Ruby, dan lain-lain. Ekstensi pada *file* Jupyter Notebook ini adalah ipynb [7][8].

E. Pandas

Pandas adalah sebuah Python *library* yang bersifat sumber terbuka atau dipublikasikan secara umum kepada orang-orang untuk analisis data. *Programmer* Python sangat membutuhkan ini untuk mempelajari dan menganalisis data untuk analisis statistik dan pengambilan keputusan. Dalam Pandas ini terdapat dua tipe data yaitu *Series* dan *DataFrame*. *DataFrame* merupakan sebuah data yang diatur dalam baris dan kolom sedangkan *Series* adalah satu kolom pada *DataFrame* [9].

F. Numpy

Numerical Python atau biasa disebut NumPy merupakan sebuah *library* Python untuk *scientific computing* terutama pada analisis data. Pada *library* ini sudah lebih baik dibandingkan *library* Python yang standar karena *library* Python yang standar lebih sederhana atau tidak memadai untuk analisis data. NumPy ini biasa digunakan pada perhitungan array multidimensi atau array dengan jumlah besar [10].

Kunci dalam tipe data NumPy yaitu array n-dimensi (ndarray). Ndarrays disebut N-dimensi karena mereka dapat memiliki sejumlah dimensi. Array satu dimensi kira-kira sama dengan *list* Python. Pada ndarray ini mirip dengan list Python tapi memiliki keunggulan yaitu teknik manipulasi data. Array ini harus bertipe homogen atau semua item pada array harus memiliki tipe yang sama [11].

Array Numpy mempunyai beberapa keunggulan dibandingkan *list* Python. Keunggulan itu berfokus pada manipulasi performa dari data yang homogen. Berikut manfaat dari keunggulan tersebut yaitu:

1. *Contiguous allocation in memory*
   Alokasi ini memberikan manfaat yang baik karena memastikan semua elemen array bisa diakses langsung di waktu tertentu sejak awal pelaksanaan array.
2. *Vectorized operation*
   Operasi vektor adalah teknik operasi di semua elemen tanpa pengkodean *loop* yang eksplisit. Operasi vektor sering kali lebih efisien dalam pengeksekusian dibandingkan dengan *loop* yang diimplementasikan dalam tingkat bahasa yang lebih tinggi. Operasi ini dapat mengefisien jumlah kode yang ditulis dan membantu meminimalkan kesalahan pengkodean.
3. *Boolean selection*
   Pemilihan *Boolean* adalah pola umum dengan NumPy dan Pandas sebagai tempat pemilihan elemen dari array didasarkan pada kriteria yang logis. Pemilihan ini dapat digunakan untuk memilih item yang cocok.
4. *Sliceability*
   Sliceability memberikan *programmer* cara yang efisien untuk menentukan elemen array dalam notasi yang sesuai. Proses *slicing* menguntungkan dengan cara memanfaatkan alokasi memori yang berdekatan dari array untuk memaksimalkan akses ke seri item [12].

G. Matplotlib

Matplotlib merupakan sebuah *library* Python yang berfokus pada visualisasi data *multiplatform* yang dibangun di atas array. Matplotlib sangat penting untuk menggunakan di berbagai sistem operasi dan *backend* grafis. Versi Matplotlib yang baru sangat mudah untuk membuat gaya plot. Matplotlib ini berfungsi membuat plot grafik [13].

Ada berbagai jenis plot yang dapat dibuat dengan mudah yaitu line, scattered, bar, box, dan radial plots. Matplotlib ini bersifat *open source* sehingga *library* ini dapat digunakan secara gratis dan bebas berkontribusi pada Matplotlib *library*.

Kelebihan lain, yaitu mudah mendapatkan dukungan *online* dari komunitas di platform dan forum. Matplotlib ini dapat digunakan di Jupyter Notebook dan mendukung berbagai visualisasi data [14].

III. METODOLOGI PENELITIAN

A. Dataset Olympics

Sebelum memulai analisa dapat memuat dataset terlebih dahulu yang diambil dari website Kaggle. Terdapat 5 dataset pada Olimpiade Tokyo 2021 dan 4 dataset pada Olimpiade tahun 1896 – 2021. Dalam dataset Olimpiade Tokyo 2021 terdapat data athlete, coaches, entriesgender, medals, dan teams. Sedangkan pada dataset Olimpiade tahun 1896 – 2021 terdapat data athletes, hosts, medals, dan results. Dataset dalam penelitian ini memiliki dua format yaitu XLSX dan CSV.

B. Mengidentifikasi Dataset

Dalam eksplorisasi penting untuk memahami dataset dengan cara mengidentifikasi dataset tersebut. Salah satu caranya adalah menggunakan fungsi info() untuk memberikan tampilan informasi secara detail pada dataframe. Dalam fungsi ini akan menampilkan jumlah baris, nama kolom, jumlah data, dan tipe data.

C. Menganalisis Dataset Athlete

Eksplorasi pada kali ini akan menganalisis dataset athlete sebagai dataset pertama pada Olimpiade Tokyo 2021. Dataset athlete ini menampilkan informasi yaitu terdapat 11.805 baris, tiga nama kolom (Name, NOC, dan Discipline), dan bertipe data object. Selain itu terdapat 11.602 nama, 206 negara, dan 46 cabang olahraga yang unik. Pada dataset ini dapat mencari informasi atlet yang unik karena terdapat informasi yang menarik setelah melihat data dari fungsi info().

D. Menganalisis Dataset Gender

Eksplorasi berikutnya akan menganalisis dataset gender sebagai dataset kedua pada Olimpiade Tokyo 2021. Dataset athlete ini menampilkan informasi yaitu terdapat 46 baris, empat nama kolom (Discipline, Female, Male, dan Total), dan bertipe data object dan int. Pada dataset ini terdapat data male dan female sehingga dapat mencari perbandingan jumlah gender.

E. Menganalisis Dataset Coach

Eksplorasi berikutnya akan menganalisis dataset coach sebagai dataset ketiga pada Olimpiade Tokyo 2021. Dataset coach ini menampilkan informasi yaitu terdapat 394 baris, empat nama kolom (Name, NOC, Discipline, dan Event), dan bertipe data object. Pada dataset ini dapat mencari informasi perbandingan jumlah coach tertinggi dan terendah karena terdapat data coach yang dapat dihitung jumlahnya dan terdapat informasi negara asal.

F. Menganalisis Dataset Medals

Eksplorasi berikutnya akan menganalisis dataset medals sebagai dataset ketiga pada Olimpiade Tokyo 2021. Dataset coach ini menampilkan informasi yaitu terdapat 93 baris, tujuh nama kolom (Rank, Team/NOC, Gold, Silver, Bronze, Total, dan Rank by Total), dan bertipe data object dan int. Pada dataset ini dapat mencari informasi total medali dengan menggunakan map. Analisis berikutnya dapat mencari perbandingan secara umum jumlah medali karena terdapat informasi data setiap medali.

G. Menganalisis Dataset Teams

Eksplorasi pada kali ini akan menganalisis dataset teams sebagai dataset pertama pada Olimpiade Tokyo 2021. Dataset athlete ini menampilkan informasi yaitu terdapat 743 baris, empat nama kolom (Name, Discipline, NOC, dan Event), dan bertipe data object. Pada dataset ini terdapat data team dan data discipline sehingga dapat mencari cabang olahraga yang diikuti team. Setelah mengetahui cabang olahraga selanjutnya dapat mencari jumlah team yang berpartisipasi dalam cabang olahraga tersebut.

H. Menganalisis Dataset Olimpiade Tahun 1896-2021

Pada eksplorasi kali ini mencari informasi dari Olimpiade sebelumnya karena akan lebih menarik untuk melanjutkan eksplorasi lebih dalam. Pertama dapat memuat 4 dataset yang ada. Dalam menampilkan dataset dapat memanggil variabel olympic_hosts, olympic_athletes, olympic_medals, dan olympic_results.

I. Menganalisis Tempat & Tahun Olimpiade

Setelah memuat dataset berikutnya dapat mengeksplor dari 4 dataset tersebut. Pertama akan mencari tempat dan tahun Olimpiade yang pernah diadakan. Ketika sudah mengetahui tempat dan tahun Olimpiade maka berikutnya akan mencari Olimpiade berdasarkan musim.

J. Menganalisis Umur Athletes

Pada eksplorasi kali ini akan menganalisis rata-rata umur atlet pada seluruh Olimpiade. Dalam mencari eksplorasi tersebut dapat mencari tahun untuk mendapatkan data umur. Setelah mendapat data tersebut selanjutnya menggabungkan nama, cabang olahraga, dan umur untuk dapat mencari umur rata-rata di setiap cabang olahraga.

IV. HASIL DAN TEMUAN

A. Atlet yang Unik

Berikut adalah informasi atlet yang unik.

```
atlet = athletes_df.groupby('Name')['NOC','Discipline'].nunique()

kewarganegaraan = atlet['NOC'] == 1
kewarganegaraan1 = atlet['NOC'] > 1
kewarganegaraan2 = atlet['NOC'] > 2

cabang_olahraga = atlet['Discipline'] == 1
cabang_olahraga1 = atlet['Discipline'] > 1
cabang_olahraga2 = atlet['Discipline'] > 2
```

Gambar 1. Mengelompokkan Data

Berikut memuat hasil atlet yang memiliki satu atau lebih kewarganegaraan dari Gambar 1.

```
print(f'Jumlah atlet dengan satu kewarganegaraan adalah: {atlet[kewarganegaraan].shape[0]} atlet')
print(f'Jumlah atlet dengan dua kewarganegaraan adalah: {atlet[kewarganegaraan1].shape[0]} atlet')
print(f'Jumlah atlet yang kewarganegaraan lebih dari dua adalah: {atlet[kewarganegaraan2].shape[0]} atlet')
```

```
Jumlah atlet dengan satu kewarganegaraan adalah: 11053 atlet
Jumlah atlet dengan dua kewarganegaraan adalah: 9 atlet
Jumlah atlet yang kewarganegaraan lebih dari dua adalah: 0 atlet
```

Gambar 2. Hasil Atlet Kewarganegaraan

Berikut memuat hasil atlet yang bertanding satu atau lebih cabang olahraga dari Gambar 1.

```
print(f'Atlet yang bertanding dalam satu cabang olahraga adalah: {atlet[cabang_olahraga].shape[0]} atlet')
print(f'Atlet yang bertanding dalam dua cabang olahraga adalah: {atlet[cabang_olahraga1].shape[0]} atlet')
print(f'Atlet yang bertanding lebih dari dua cabang olahraga adalah: {atlet[cabang_olahraga2].shape[0]} atlet')
```

```
Atlet yang bertanding dalam satu cabang olahraga adalah: 11041 atlet
Atlet yang bertanding dalam dua cabang olahraga adalah: 21 atlet
Atlet yang bertanding lebih dari dua cabang olahraga adalah: 0 atlet
```

Gambar 3. Hasil Atlet Cabang Olahraga

Dari data di atas terlihat bahwa 9 atlet yang memiliki dua kewarganegaraan dan 21 atlet mengikuti dua cabang olahraga. Berikut adalah nama atlet yang memiliki dua kewarganegaraan. Kesamaan nama tidak otomatis merupakan atlet yang memiliki dua kewarganegaraan seperti Portela Teresa. Portela Teresa adalah nama atlet yang berbeda yang berasal dari Portugis dan Spanyol. Berikut adalah nama atlet yang memiliki 2 kewarganegaraan dan 2 cabang olahraga.

```
atlet_2k = list(set(atlet[kewarganegaraan1].index))
idxs = np.where(athletes_df['Name'].apply(lambda x: x in atlet_2k))[0]
df_atlet_2k = athletes_df.iloc[idxs]
df_atlet_2k.set_index(['Name'],inplace=True)
df_atlet_2k
```

Name	NOC	Discipline
ALVAREZ Jorge	Cuba	Shooting
ALVAREZ Jorge	Honduras	Football
HALL James	Great Britain	Artistic Gymnastics
HALL James	United States of America	Shooting
KURBANOV Ruslan	Kazakhstan	Fencing
KURBANOV Ruslan	Uzbekistan	Athletics
LI Qian	People's Republic of China	Boxing
LI Qian	Poland	Table Tennis
MARTIN Daniel	Ireland	Cycling Road
MARTIN Daniel	Romania	Swimming
PEREZ Maria	Puerto Rico	Judo
PEREZ Maria	Spain	Athletics
PEREZ Paola	Ecuador	Athletics
PEREZ Paola	Venezuela	Marathon Swimming
PORTELA Teresa	Portugal	Canoe Sprint
PORTELA Teresa	Spain	Canoe Sprint
WANG Yang	People's Republic of China	Sailing
WANG Yang	Slovakia	Table Tennis

Gambar 4. Nama Atlet yang Memiliki Dua Kewarganegaraan

```
atlet_2co = list(set(atlet[cabang_olahraga1].index))
idxs = np.where(athletes_df['Name'].apply(lambda x: x in atlet_2co))[0]
df_atlet_2k = athletes_df.iloc[idxs]
df_atlet_2k.set_index(['Name'],inplace=True)
df_atlet_2k
```

Name	NOC	Discipline
ALVAREZ Jorge	Cuba	Shooting
ALVAREZ Jorge	Honduras	Football
CHEN Yang	People's Republic of China	Athletics
CHEN Yang	People's Republic of China	Hockey
DYGERT Chloe	United States of America	Cycling Road
DYGERT Chloe	United States of America	Cycling Track
GANNA Filippo	Italy	Cycling Road
GANNA Filippo	Italy	Cycling Track
HALL James	Great Britain	Artistic Gymnastics
HALL James	United States of America	Shooting
HAVIK Yoeri	Netherlands	Cycling Road
HAVIK Yoeri	Netherlands	Cycling Track
KIM Hyunsoo	Republic of Korea	Baseball/Softball
KIM Hyunsoo	Republic of Korea	Rugby Sevens
KOPECKY Lotte	Belgium	Cycling Road
KOPECKY Lotte	Belgium	Cycling Track
KOVACS Zsofia	Hungary	Artistic Gymnastics
KOVACS Zsofia	Hungary	Triathlon
KURBANOV Ruslan	Kazakhstan	Fencing
KURBANOV Ruslan	Uzbekistan	Athletics
LI Qian	People's Republic of China	Boxing
LI Qian	Poland	Table Tennis
MARTIN Daniel	Ireland	Cycling Road
MARTIN Daniel	Romania	Swimming
PALTRINIERI Gregorio	Italy	Marathon Swimming
PALTRINIERI Gregorio	Italy	Swimming
PEREZ Maria	Puerto Rico	Judo
PEREZ Maria	Spain	Athletics
PEREZ Paola	Ecuador	Athletics
PEREZ Paola	Venezuela	Marathon Swimming
SUN Jiajun	People's Republic of China	Cycling Road
SUN Jiajun	People's Republic of China	Swimming
van ROUWENDAAL Sharon	Netherlands	Marathon Swimming
van ROUWENDAAL Sharon	Netherlands	Swimming
WANG Yang	People's Republic of China	Sailing
WANG Yang	Slovakia	Table Tennis
WATANABE Yuta	Japan	Badminton
WATANABE Yuta	Japan	Basketball
WELLBROCK Florian	Germany	Marathon Swimming
WELLBROCK Florian	Germany	Swimming
ZHANG Xin	People's Republic of China	Football
ZHANG Xin	People's Republic of China	Skateboarding

Gambar 5. Nama Atlet yang Mengikuti Dua Cabang Olahraga

B. Perbandingan Jumlah Gender

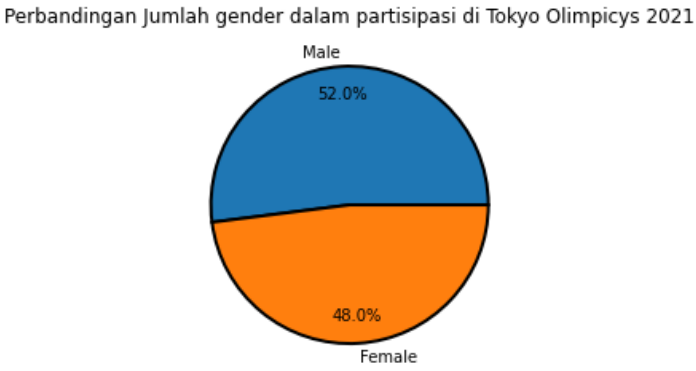
Berikut adalah bentuk visualisasi pie chart yang berisi perbandingan gender.

```
m = entriesGender_df['Male'].sum()
f = entriesGender_df['Female'].sum()
labl = ["Male","Female"]
total = [m,f]

plt.pie(total,
        labels=labl,
        autopct='%1.1f%%',
        pctdistance=0.80,
        wedgeprops={'edgecolor':'black', 'linewidth': '2'})
plt.title('Perbandingan Jumlah gender dalam partisipasi di Tokyo Olimpicys 2021')

fig = plt.gcf()

plt.show()
```



Gambar 6. Pie Chart Perbandingan Jumlah Gender

Dari visualiasi pie chart Gambar 6 terdapat bahwa partisipasi laki-laki lebih dominan dibandingkan perempuan yang hanya 48% saja. Perbandingan jumlah gender dalam cabang olahraga dapat dieksplorasi berdasarkan Gambar 6.

C. Negara Dengan Jumlah Coach Terbanyak dan Terendah

Berikut analisa menghitung jumlah coach terbanyak dalam setiap negara.

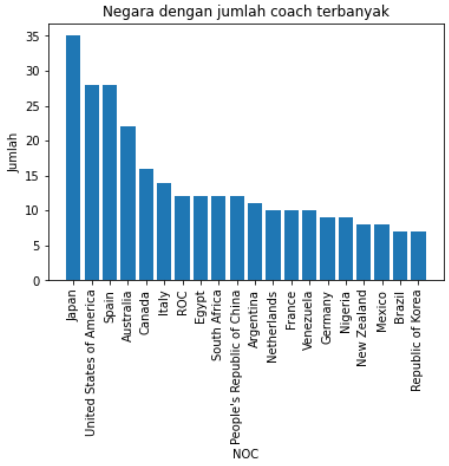
```
jumlah_coach_terbanyak = coaches_df.groupby('NOC').size()
jumlah_coach_terbanyak = jumlah_coach_terbanyak.to_frame('Count')
jumlah_coach_terbanyak = jumlah_coach_terbanyak.sort_values('Count', ascending=False).head(20)
jumlah_coach_terbanyak.head(20)
```

NOC	Count
Japan	35
United States of America	28
Spain	28
Australia	22
Canada	16
Italy	14
ROC	12
Egypt	12
South Africa	12
People's Republic of China	12
Argentina	11
Netherlands	10
France	10
Venezuela	10
Germany	9
Nigeria	9
New Zealand	8
Mexico	8
Brazil	7
Republic of Korea	7

Gambar 7. Analisa Jumlah 20 Coach Terbanyak

```
jumlah_coach_terbanyak = coaches_df.groupby('NOC').size()
jumlah_coach_terbanyak = jumlah_coach_terbanyak.to_frame('Count')
jumlah_coach_terbanyak = jumlah_coach_terbanyak.sort_values('Count', ascending=False).head(20)

plt.bar(jumlah_coach_terbanyak.index.values, jumlah_coach_terbanyak['Count'])
plt.title('Negara dengan jumlah coach terbanyak')
plt.xlabel('NOC')
plt.ylabel('Jumlah')
plt.xticks(rotation=90)
plt.show()
```



Gambar 8. Bar Chart Jumlah Coach Terbanyak

Dari data di atas terlihat bahwa Jepang memiliki coach terbanyak dengan jumlah 35 coach. Di posisi kedua adalah Amerika dengan jumlah 28 coach dan urutan ketiga yaitu Spanyol dengan jumlah coach sama dengan Amerika yaitu 28 coach. Berikut akan mencari jumlah coach terendah.

Berikut analisa menghitung jumlah coach terendah dalam setiap negara.

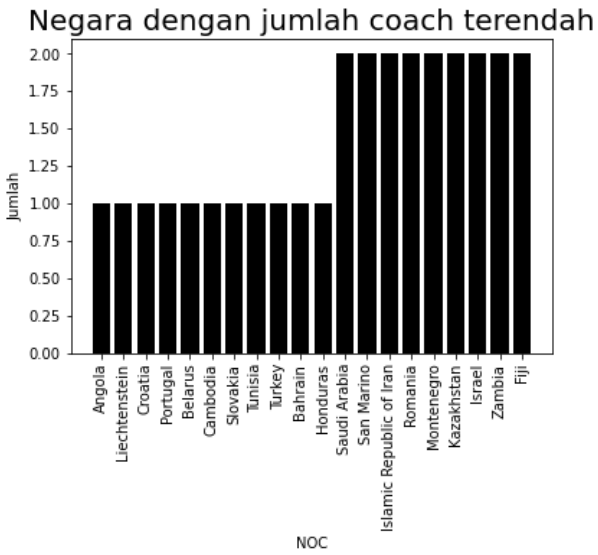
```
jumlah_coach_terendah = coaches_df.groupby('NOC').size()
jumlah_coach_terendah = jumlah_coach_terendah.to_frame('Count')
jumlah_coach_terendah = jumlah_coach_terendah.sort_values('Count')
jumlah_coach_terendah.head(20)
```

NOC	Count
Angola	1
Liechtenstein	1
Croatia	1
Portugal	1
Belarus	1
Cambodia	1
Slovakia	1
Tunisia	1
Turkey	1
Bahrain	1
Honduras	1
Saudi Arabia	2
San Marino	2
Islamic Republic of Iran	2
Romania	2
Montenegro	2
Kazakhstan	2
Israel	2
Zambia	2
Fiji	2

Gambar 9. Analisa Jumlah Coach Terendah

```
jumlah_coach_terendah = coaches_df.groupby('NOC').size()
jumlah_coach_terendah = jumlah_coach_terendah.to_frame('Count')
jumlah_coach_terendah = jumlah_coach_terendah.sort_values('Count').head(20)

plt.bar(jumlah_coach_terendah.index.values, jumlah_coach_terendah['Count'], color='Black')
plt.title('Negara dengan jumlah coach terendah', fontsize=20)
plt.xlabel('NOC')
plt.ylabel('Jumlah')
plt.xticks(rotation=90)
plt.show()
```



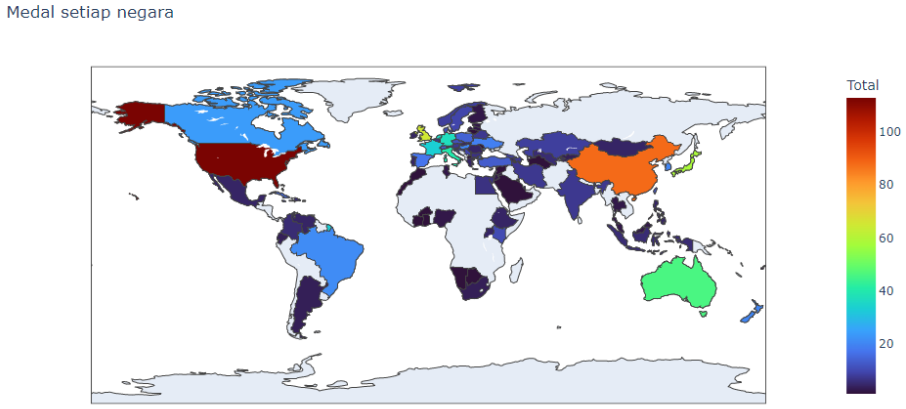
Gambar 10. Bar Chart Jumlah Coach Terendah

Dari data di atas terlihat bahwa ada 11 negara dengan coach terendah dan salah satunya adalah negara Angola yang memiliki 1 coach. Sedang negara yang memiliki 2 coach contohnya adalah Saudi Arabia.

D. Membuat Map Jumlah Medali dan Perbandingan Total Medali

Berikut akan mencari jumlah medali yang divisualisasikan menggunakan map. Pertama, dapat membuat visualisasi map yang berisi jumlah medali pada setiap negara.

```
px.choropleth(medals_df, locations="Team/NOC",
                    locationmode='country names',
                    color="Total",
                    hover_name="Total",
                    color_continuous_scale="turbo",
                    title='Medal setiap negara')
```



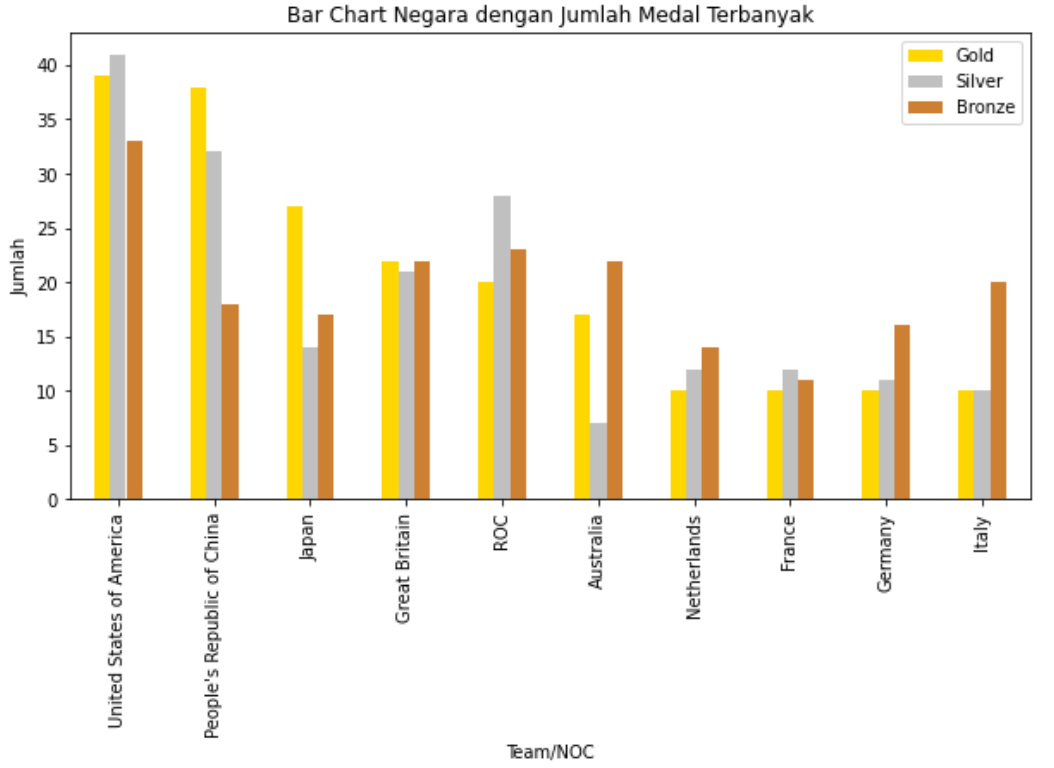
Gambar 11. Map Jumlah Medali Setiap Negara

Dari map di atas negara dengan warna merah merupakan total medal yang paling banyak contohnya seperti negara Amerika sedangkan negara dengan warna biru tua merupakan total medal yang paling sedikit salah satu contohnya adalah Saudi Arabia.

Berikut visualisasi bar chart pada perbandingan total medali gold, silver, dan bronze pada 10 negara.

```
medals_df[:10].plot(x='Team/NOC',
           y=['Gold', 'Silver', 'Bronze'],
           color=['#FFD700', '#C0C0C0', '#CD7F32'],
           kind='bar', figsize=(10,5))

plt.title('Bar Chart Negara dengan Jumlah Medal Terbanyak')
plt.ylabel('Jumlah')
plt.show()
```



Gambar 12. Bar Chart Perbandingan Total Medali

Data di atas terlihat beberapa negara dengan mendapatkan medali yang sudah didapat tetapi Gambar 12 masih secara singkat menggambarkan medali yang diraih dari setiap negara. Data pada Gambar 12 terlihat bahwa Amerika mendominasi medali *gold*, *silver*, dan *bronze*.

E. Cabang Olahraga yang Diikuti Team dan Jumlah Team yang Bepartisipasi

Berikut adalah analisa cabang olahraga yang diikuti oleh team.

```
1 teams_df['Discipline'].unique()
```

```
array(['3x3 Basketball', 'Archery', 'Artistic Gymnastics',
       'Artistic Swimming', 'Athletics', 'Baseball/Softball',
       'Basketball', 'Beach Volleyball', 'Cycling Track', 'Fencing',
       'Football', 'Handball', 'Hockey', 'Rhythmic Gymnastics',
       'Rugby Sevens', 'Swimming', 'Table Tennis', 'Triathlon',
       'Volleyball', 'Water Polo'], dtype=object)
```

Gambar 13. Analisa Cabang Olahraga yang Diikuti Oleh Team

Berikut adalah analisa jumlah team yang berpartisipasi dalam setiap cabang olahraga.

```
1 teams_df.groupby('Discipline')['Name'].count().sort_values(ascending=False)
```

```
Discipline
Swimming              113
Athletics              79
Cycling Track          64
Archery                53
Fencing                52
Beach Volleyball       48
Table Tennis           48
Artistic Swimming      32
Football               28
Artistic Gymnastics    24
Volleyball             24
Basketball             24
Handball               24
Hockey                 24
Rugby Sevens           24
Water Polo             22
Triathlon              18
3x3 Basketball         16
Rhythmic Gymnastics    14
Baseball/Softball      12
Name: Name, dtype: int64
```

Gambar 14. Analisa Jumlah Team dalam Cabang Olahraga

Data dari **Error! Reference source not found.** terlihat bahwa *swimming* merupakan partisipasi yang paling diminati banyak team. Berikutnya akan mencari negara dengan jumlah team yang terbanyak.

F. Tempat dan Tahun Olimpiade Berdasarkan Musim

Berikut akan mencari seluruh tempat & tahun Olimpiade yang pernah diadakan.

```
1  olympic_results['slug_game'].unique()
```

```
array(['tokyo-2020', 'pyeongchang-2018', 'rio-2016', 'sochi-2014',
       'london-2012', 'vancouver-2010', 'beijing-2008', 'turin-2006',
       'athens-2004', 'salt-lake-city-2002', 'sydney-2000', 'nagano-1998',
       'atlanta-1996', 'lillehammer-1994', 'barcelona-1992',
       'albertville-1992', 'seoul-1988', 'calgary-1988',
       'los-angeles-1984', 'sarajevo-1984', 'moscow-1980',
       'lake-placid-1980', 'montreal-1976', 'innsbruck-1976',
       'munich-1972', 'sapporo-1972', 'mexico-city-1968', 'grenoble-1968',
       'tokyo-1964', 'innsbruck-1964', 'rome-1960', 'squaw-valley-1960',
       'melbourne-1956', 'cortina-d-ampezzo-1956', 'helsinki-1952',
       'oslo-1952', 'london-1948', 'st-moritz-1948', 'berlin-1936',
       'garmisch-partenkirchen-1936', 'los-angeles-1932',
       'lake-placid-1932', 'amsterdam-1928', 'st-moritz-1928',
       'paris-1924', 'chamonix-1924', 'antwerp-1920', 'stockholm-1912',
       'london-1908', 'st-louis-1904', 'paris-1900', 'athens-1896'],
      dtype=object)
```

Gambar 15. Analisa Tempat & Tahun Olimpiade

Berikut hasil analisa tempat dan tahun olimpiade pada saat musim panas.

```
1 tempat_tahun_olimpiade_summer = olympic_results[(olympic_results['game_season'] == 'Summer')]
2 jumlah_tempat_tahun_olimpiade_summer = tempat_tahun_olimpiade_summer['slug_game'].unique()
3 jumlah_tempat_tahun_olimpiade_summer = pd.DataFrame(jumlah_tempat_tahun_olimpiade_summer, columns = ['Season'])
4 jumlah_tempat_tahun_olimpiade_summer
```

	Season
0	tokyo-2020
1	rio-2016
2	london-2012
3	beijing-2008
4	athens-2004
5	sydney-2000
6	atlanta-1996
7	barcelona-1992
8	seoul-1988
9	los-angeles-1984
10	moscow-1980
11	montreal-1976
12	munich-1972
13	mexico-city-1968
14	tokyo-1964
15	rome-1960
16	melbourne-1956
17	helsinki-1952
18	london-1948
19	berlin-1936
20	los-angeles-1932
21	amsterdam-1928
22	paris-1924
23	antwerp-1920
24	stockholm-1912
25	london-1908
26	st-louis-1904
27	paris-1900
28	athens-1896

Gambar 16. Analisa Olimpiade Musim Panas

Dari data di atas terlihat bahwa terdapat 29 Olimpiade yang pernah diadakan pada musim panas. Pertama kali diadakan yaitu di Athens pada tahun 1896 sedangkan terakhir diadakan di Tokyo pada tahun 2020.

Berikut hasil analisa tempat dan tahun Olimpiade pada saat musim dingin.

```
tempat_tahun_olimpiade_winter = olympic_results[(olympic_results['game_season'] == 'Winter')]
jumlah_tempat_tahun_olimpiade_winter = tempat_tahun_olimpiade_winter['slug_game'].unique()
jumlah_tempat_tahun_olimpiade_winter = pd.DataFrame(jumlah_tempat_tahun_olimpiade_winter, columns = ['Season'])
jumlah_tempat_tahun_olimpiade_winter
```

	Season
0	pyeongchang-2018
1	sochi-2014
2	vancouver-2010
3	turin-2006
4	salt-lake-city-2002
5	nagano-1998
6	lillehammer-1994
7	albertville-1992
8	calgary-1988
9	sarajevo-1984
10	lake-placid-1980
11	innsbruck-1976
12	sapporo-1972
13	grenoble-1968
14	innsbruck-1964
15	squaw-valley-1960
16	cortina-d-ampezzo-1956
17	oslo-1952
18	st-moritz-1948
19	garmisch-partenkirchen-1936
20	lake-placid-1932
21	st-moritz-1928
22	chamonix-1924

Gambar 17. Analisa Olimpiade Musim Dingin

Dari Gambar 17 terlihat bahwa terdapat 23 Olimpiade yang pernah diadakan pada musim dingin. Pertama kali diadakan yaitu di Chamonix pada tahun 1924 sedangkan terakhir diadakan di Pyeongchang pada tahun 2018.

G. Umur Athletes

Berikut akan mencari rata-rata umur dalam cabang olahraga.

```
game_pertama = olympic_athletes.first_game.str.split('\s+').str[-1].to_frame('tahun')
game_pertama1 = game_pertama.tahun.fillna('0', inplace=True)
game_pertama2 = game_pertama.tahun.astype(int).to_frame('tahun')
game_pertama2
```

	tahun
0	2020
1	2020
2	2020
3	2020
4	2016
...	...
74726	1976
74727	1976
74728	1976
74729	1976
74730	1976

74731 rows × 1 columns

Gambar 18. Mencari Data Tahun

Gambar 18 adalah mencari data tahun untuk melakukan pencarian data umur.

```
umur_atlet = (game_pertama2.tahun - olympic_athletes.athlete_year_birth).to_frame('umur')
umur_atlet1 = umur_atlet.umur.fillna(0, inplace=True)
umur_atlet1 = umur_atlet.umur.astype(int).to_frame('umur')
umur_atlet1[umur_atlet1 < 0] = 0
umur_atlet1
```

	umur
0	32
1	27
2	23
3	26
4	20
...	...
74726	21
74727	19
74728	19
74729	22
74730	16

74731 rows × 1 columns

Gambar 19. Mencari Data Umur

Gambar 19 adalah mencari data umur untuk mengetahui umur atlet. Data ini akan digabungkan dengan nama atlet sesuai urutan dari dataset Olympic_Athletes.

```
nama_athlete = olympic_athletes.athlete_full_name
data_athlete = pd.merge(nama_athlete, umur_atlet1, left_index= True, right_index = True)
data_athlete = data_athlete[data_athlete.umur != 0]
data_athlete
```

	athlete_full_name	umur
0	Dongqi CHEN	32
1	Sagen MADDALENA	27
2	Maria MARTYNOVA	23
3	Yulia KARIMOVA	26
4	Najmeh KHEDMATI	20
...	...	...
74726	Douglas WEIGLE	21
74727	Stefania BERTELE	19
74728	Walter CECCONI	19
74729	Susan KELLEY	22
74730	Andrew STROUKOFF	16

72150 rows × 2 columns

Gambar 20. Data Nama Atlet dan Umur

Gambar 20 adalah data nama atlet dan umur atlet pada seluruh Olimpiade. Data ini akan digabungkan dengan cabang olahraga sesuai nama atlet.

```
nama_athlete1 = olympic_results.athlete_full_name
cabang_olahraga = olympic_results.discipline_title
data_athlete1 = pd.merge(nama_athlete1, cabang_olahraga, left_index= True, right_index = True)
data_athlete1
```

	athlete_full_name	discipline_title
0	Fatima GALVEZ	Shooting
1	Alberto FERNANDEZ	Shooting
2	Alessandra PERILLI	Shooting
3	Gian Marco BERTI	Shooting
4	Madelynn Ann BERNAU	Shooting
...	...	...
21305	Viggo JENSEN	Weightlifting
21306	Alexandros Nikolopoulos	Weightlifting
21307	Viggo JENSEN	Weightlifting
21308	Launceston ELLIOT	Weightlifting
21309	Sotirios VERSIS	Weightlifting

21310 rows × 2 columns

Gambar 21. Data Nama Atlet dan Cabang Olahraga

Gambar 21 merupakan data nama atlet dan cabang olahraga. Data ini akan digabungkan dengan data umur.

```
data_athlete2 = pd.merge(data_athlete, data_athlete1, how='inner', on='athlete_full_name')
data_athlete2
```

	athlete_full_name	umur	discipline_title
0	Yulia KARIMOVA	26	Shooting
1	Nina CHRISTEN	22	Shooting
2	Nina CHRISTEN	22	Shooting
3	Mary Carolynn TUCKER	19	Shooting
4	Eglys CRUZ	24	Shooting
...	...	...	...
16677	Clarence Olivier GAMBLE	23	Tennis
16678	Arthur Yancey WEAR	24	Tennis
16679	Dimitrios LOUNDRAS	11	Gymnastics Artistic
16680	Lyudmila PAKHOMOVA	30	Figure skating
16681	Colleen O'CONNOR	25	Figure skating

16682 rows × 3 columns

Gambar 22. Data Informasi Atlet

Gambar 22 adalah data gabungan dari nama atlet, umur, dan cabang olahraga. Berikutnya, mencari rata-rata umur tertua dan termuda pada cabang olahraga.

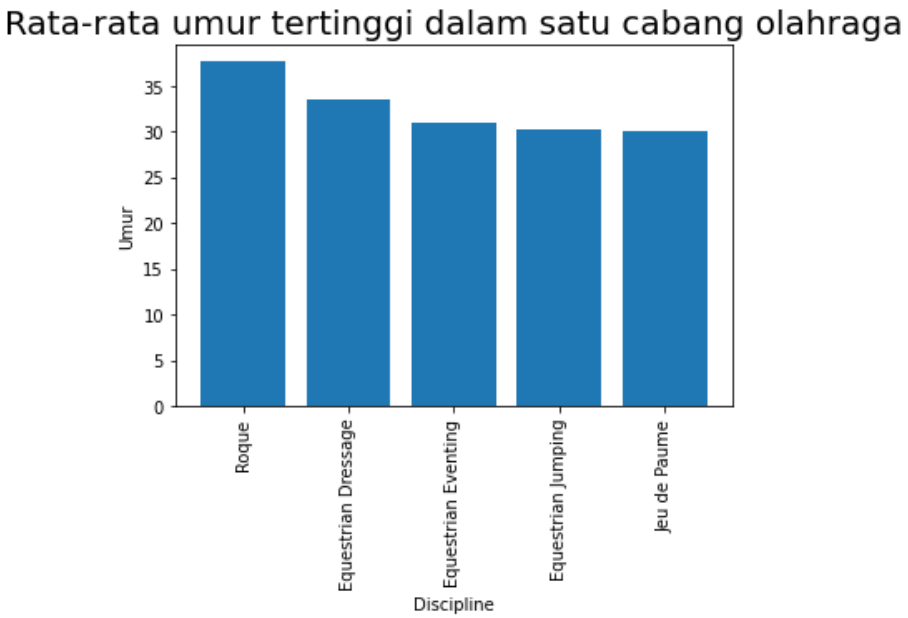
```
data_athlete3 = data_athlete2.groupby('discipline_title').mean().sort_values('umur', ascending=False).head()
data_athlete3
```

discipline_title	umur
Roque	37.750000
Equestrian Dressage	33.492308
Equestrian Eventing	30.954545
Equestrian Jumping	30.260870
Jeu de Paume	30.000000

Gambar 23. Rata-Rata Umur Tertua

Gambar 23 adalah rata-data umur tertua pada cabang olahraga. Cabang olahraga roque merupakan cabang olahraga dengan rata-rata atlet tertua dengan rata-rata umur 37.75.

```
plt.bar(data_athlete3.index.values, data_athlete3['umur'])
plt.title('Rata-rata umur tertinggi dalam satu cabang olahraga', fontsize=20)
plt.xlabel('Discipline')
plt.ylabel('Umur')
plt.xticks(rotation=90)
plt.show()
```



Gambar 24. Visualisasi Rata-Rata Umur Tertua

Gambar 24 adalah visualisasi dari rata-rata umur tertua. Posisi pertama terdapat cabang olahraga roque dengan rata-rata 37.75. Posisi kedua yaitu equestrian dressage dengan rata-rata umur 33.5. Posisi ketiga yaitu equestrian eventing dengan rata-rata umur 31.

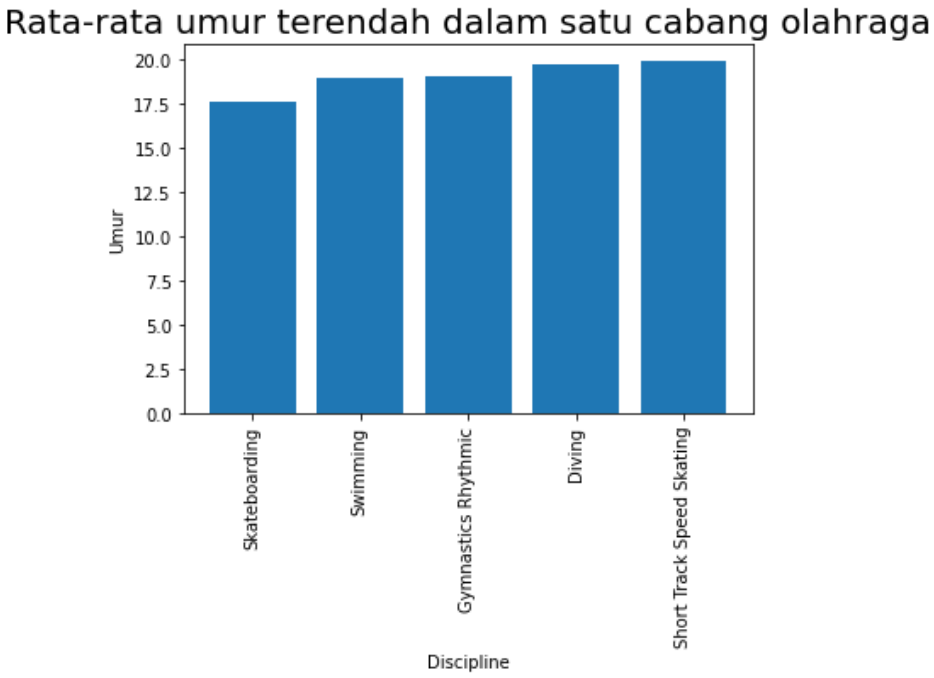
```
data_athlete4 = data_athlete2.groupby('discipline_title').mean().sort_values('umur', ascending=True).head()
data_athlete4
```

discipline_title	umur
Skateboarding	17.666667
Swimming	19.005722
Gymnastics Rhythmic	19.062500
Diving	19.794760
Short Track Speed Skating	19.934066

Gambar 25. Rata-Rata Umur Termuda

Gambar 25 adalah rata-data umur termuda pada cabang olahraga. Cabang olahraga skateboarding merupakan cabang olahraga dengan rata-rata atlet tertua dengan rata-rata umur 17.666667.

```
plt.bar(data_athlete4.index.values, data_athlete4['umur'])
plt.title('Rata-rata umur terendah dalam satu cabang olahraga', fontsize=20)
plt.xlabel('Discipline')
plt.ylabel('Umur')
plt.xticks(rotation=90)
plt.show()
```



Gambar 26. Visualisasi Rata-Rata Umur Termuda

Gambar 26 adalah visualisasi dari rata-rata umur termuda. Posisi pertama terdapat cabang olahraga skateboarding dengan rata-rata 17.7. Posisi kedua yaitu swimming dengan rata-rata umur 19. Posisi ketiga yaitu gymnastics rhythmic dengan rata-rata umur 19.1.

V. KESIMPULAN

Simpulan eksplorasi yang sudah dibuat dan analisis data Olimpiade dengan menggunakan dataset “2021 Olympics in Tokyo” dan “Olympic Games, 1986-2021” yang berasal dari Kaggle yaitu sebagai berikut:

1. Dataset Olimpiade Tokyo 2021 yang terdiri dari data Athletes, Coaches, EntriesGender, Medals, dan Teams dan dataset Olimpiade 1896-2021 yang terdiri dari data Olympic_athletes, Olympic_hosts, Olympic_medals, dan Olympic_results dapat diolah menjadi sekumpulan informasi yang menarik seperti membuat tabel yang berisi data.

2. Dari dataset Olimpiade yang sebelumnya diolah seperti tabel, dapat divisualisasikan dengan memakai matplotlib seperti pie plot, bar plot dan scatter plot untuk memudahkan membaca informasi yang disampaikan.
3. Eksplorasi dari dataset Olimpiade menghasilkan informasi data. Berikut beberapa contoh eksplorasi:
   - Atlet yang unik yang memiliki lebih dari satu kewarganegaraan dan mengikuti lebih dari satu cabang olahraga.
   - Perbedaan cabang olahraga musim panas dan dingin.
   - Atlet yang menjadi the GOAT di cabang olahraga yang selalu ada.

DAFTAR PUSTAKA

[1] G. Herman, "What Are the Summer Olympics?" 2016.
[2] P. P. Anzari, N. P. Fariza, U. N. Malang, and J. Surabaya, "Jurnal kajian media 2021," vol. 5, no. 1, pp. 39–49, 2021.
[3] V. Gallego, H. Nishiura, R. Sah, and A. J. Rodriguez-Morales, "The COVID-19 outbreak and implications for the Tokyo 2020 Summer Olympic Games," Travel Medicine and Infectious Disease, vol. 34, no. January, 2020, doi: 10.1016/j.tmaid.2020.101604.
[4] A. S. Bahri et al., PENGANTAR PENELITIAN PENDIDIKAN (Sebuah Tinjauan Teori dan Praktis). Widina Bhakti Persada Bandung, 2021.
[5] E. S. Hamid and Y. S. Susilo, "Strategi Pengembangan Usaha Mikro Kecil Dan Menengah Di Provinsi Daerah Istimewa Yogyakarta*," Jurnal Ekonomi Pembangunan: Kajian Masalah Ekonomi dan Pembangunan, vol. 12, no. 1, p. 45, 2015, doi: 10.23917/jep.v12i1.204.
[6] Alvaro Fuentes, "Become a Python Data Analyst," 2018.
[7] J. Cook, Docker for Data Science. 2017. doi: 10.1007/978-1-4842-3012-1.
[8] Marc Wintjen, Practical Data Analysis Using Jupyter Notebook: Learn how to speak the language of data by extracting useful and actionable insights using Python. 2020.
[9] P. Dataframe, C. Chapter, and T. Series, "Daniel Y. Chen-Pandas for Everyone. Python Data Analysis-Addison-Wesley Professional (2017)".
[10] F. Nelli, Python Data Analytics. 2018. doi: 10.1007/978-1-4842-3913-1.
[11] J. Nunez-Iglesias, S. van der Walt, and H. Dashnow, Elegant SciPy.
[12] M. Heydt, Learning pandas, vol. 66. 2012.
[13] J. VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, 2016.
[14] A. Yim, C. Chung, and A. Yu, Matplotlib for Python Developers: Effective techniques for data visualization with Python, 2nd Edition. Packt Publishing, 2018.