Predictive Model for Cricketers' Scores in Test Matches

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")

#Read File
df=pd.read_csv("cricketers.csv")

df
```

```
      Unnamed: 0.1  Unnamed: 0               Name Date_Of_Birth
Country  \
0                0           2   Aaftab Alam Khan    31/01/1986
Malta
1                1           3       Aamer Hameed    18/10/1954
Pakistan
2                2           4        Aamer Hanif    04/10/1971
Pakistan
3                3           5        Aamer Ikram    07/07/1979
Portugal
4                4           6        Aamer Jamal    05/07/1996
Pakistan
...            ...         ...                ...           ...
...
6805          6830        6857  Zulqarnain Haider    20/01/1989
Spain
6806          6831        6858  Zulqarnain Haider    03/04/1986
Pakistan
6807          6832        6859        Zulquarnain    25/05/1962
Pakistan
6808          6833        6860             L Zulu    14/10/1992
Eswatini
6809          6834        6861             P Zuze    18/09/2006
Malawi

      Test   ODI   T20
0      NaN   NaN  27.0
1      NaN   2.0   NaN
2      NaN   5.0   NaN
3      NaN   NaN   4.0
4      2.0   NaN   4.0
...    ...   ...   ...
6805   NaN   NaN  11.0
6806   1.0   4.0   3.0
6807   3.0  16.0   NaN
6808   NaN   NaN   6.0
6809   NaN   NaN   4.0
```

```
[6810 rows x 8 columns]
```

```
# Display the first few rows of the DataFrame
df.head()
```

```
   Unnamed: 0.1  Unnamed: 0              Name Date_Of_Birth   Country
Test  \
0             0           2  Aaftab Alam Khan    31/01/1986     Malta
NaN
1             1           3      Aamer Hameed    18/10/1954  Pakistan
NaN
2             2           4       Aamer Hanif    04/10/1971  Pakistan
NaN
3             3           5       Aamer Ikram    07/07/1979  Portugal
NaN
4             4           6       Aamer Jamal    05/07/1996  Pakistan
2.0

   ODI   T20
0  NaN  27.0
1  2.0   NaN
2  5.0   NaN
3  NaN   4.0
4  NaN   4.0
```

```
#Display the first 10 rows of the DataFrame
df.head(10)
```

```
   Unnamed: 0.1  Unnamed: 0              Name Date_Of_Birth
Country  \
0             0           2  Aaftab Alam Khan    31/01/1986
Malta
1             1           3      Aamer Hameed    18/10/1954
Pakistan
2             2           4       Aamer Hanif    04/10/1971
Pakistan
3             3           5       Aamer Ikram    07/07/1979
Portugal
4             4           6       Aamer Jamal    05/07/1996
Pakistan
5             5           7       Aamer Malik    03/01/1963
Pakistan
6             6           8       Aamer Yamin    26/06/1990
Pakistan
7             7           9      Aamir Kaleem    20/11/1981
Oman
8             8          10         Aamir Lal    10/08/1990  South
Korea
9             9          11       Aamir Nazir    02/01/1971
```

```
Pakistan
```

```
   Test   ODI   T20
0   NaN   NaN  27.0
1   NaN   2.0   NaN
2   NaN   5.0   NaN
3   NaN   NaN   4.0
4   2.0   NaN   4.0
5  14.0  24.0   NaN
6   NaN   4.0   2.0
7   NaN   3.0  33.0
8   NaN   NaN   4.0
9   6.0   9.0   NaN
```

```
#Display the last few rows of the DataFrame
df.tail()
```

```
      Unnamed: 0.1  Unnamed: 0               Name Date_Of_Birth
Country  \
6805          6830        6857  Zulqarnain Haider    20/01/1989
Spain
6806          6831        6858  Zulqarnain Haider    03/04/1986
Pakistan
6807          6832        6859        Zulquarnain    25/05/1962
Pakistan
6808          6833        6860             L Zulu    14/10/1992
Eswatini
6809          6834        6861             P Zuze    18/09/2006
Malawi
```

```
      Test   ODI   T20
6805   NaN   NaN  11.0
6806   1.0   4.0   3.0
6807   3.0  16.0   NaN
6808   NaN   NaN   6.0
6809   NaN   NaN   4.0
```

```
#To check columns of DataFrame(show all column names)
df.columns
```

```
Index(['Unnamed: 0.1', 'Unnamed: 0', 'Name', 'Date_Of_Birth',
'Country',
       'Test', 'ODI', 'T20'],
      dtype='object')
```

```
#Check for missing Data/null values
df.isnull().sum()
```

```
Unnamed: 0.1        0
Unnamed: 0          0
Name                0
```

```
Date_Of_Birth      69
Country             0
Test             3665
ODI              3899
T20              3095
dtype: int64
```

```
#summary of a DataFrame
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6810 entries, 0 to 6809
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Unnamed: 0.1   6810 non-null   int64  
 1   Unnamed: 0     6810 non-null   int64  
 2   Name           6810 non-null   object 
 3   Date_Of_Birth  6741 non-null   object 
 4   Country        6810 non-null   object 
 5   Test           3145 non-null   float64
 6   ODI            2911 non-null   float64
 7   T20            3715 non-null   float64
dtypes: float64(3), int64(2), object(3)
memory usage: 425.8+ KB
```

```
#To delete column from DataFrame
df.drop(['Unnamed: 0.1'],axis=1,inplace=True)
```

```
df
```

```
      Unnamed: 0               Name Date_Of_Birth   Country  Test
ODI   T20
0              2   Aaftab Alam Khan    31/01/1986     Malta   NaN
NaN  27.0
1              3       Aamer Hameed    18/10/1954  Pakistan   NaN
2.0   NaN
2              4        Aamer Hanif    04/10/1971  Pakistan   NaN
5.0   NaN
3              5        Aamer Ikram    07/07/1979  Portugal   NaN
NaN   4.0
4              6        Aamer Jamal    05/07/1996  Pakistan   2.0
NaN   4.0
...          ...                ...           ...       ...   ...   ..
.   ...
6805        6857  Zulqarnain Haider    20/01/1989     Spain   NaN
NaN  11.0
6806        6858  Zulqarnain Haider    03/04/1986  Pakistan   1.0
4.0   3.0
6807        6859        Zulquarnain    25/05/1962  Pakistan   3.0
```

```
16.0   NaN
6808        6860              L Zulu    14/10/1992  Eswatini   NaN
NaN   6.0
6809        6861              P Zuze    18/09/2006    Malawi   NaN
NaN   4.0

[6810 rows x 7 columns]
```

```
#To delete column from DataFrame
df.drop(['Unnamed: 0'],axis=1,inplace=True)
```

```
df
```

```
                    Name Date_Of_Birth   Country  Test   ODI   T20
0      Aaftab Alam Khan    31/01/1986     Malta   NaN   NaN  27.0
1          Aamer Hameed    18/10/1954  Pakistan   NaN   2.0   NaN
2           Aamer Hanif    04/10/1971  Pakistan   NaN   5.0   NaN
3           Aamer Ikram    07/07/1979  Portugal   NaN   NaN   4.0
4           Aamer Jamal    05/07/1996  Pakistan   2.0   NaN   4.0
...                 ...           ...       ...   ...   ...   ...
6805  Zulqarnain Haider    20/01/1989     Spain   NaN   NaN  11.0
6806  Zulqarnain Haider    03/04/1986  Pakistan   1.0   4.0   3.0
6807        Zulquarnain    25/05/1962  Pakistan   3.0  16.0   NaN
6808             L Zulu    14/10/1992  Eswatini   NaN   NaN   6.0
6809             P Zuze    18/09/2006    Malawi   NaN   NaN   4.0

[6810 rows x 6 columns]
```

```
#Removing Missing Values for Data Cleaning and Analysis
df.dropna(inplace=True)
```

```
#Check for missing Data/null values
df.isnull().sum()
```

```
Name             0
Date_Of_Birth    0
Country          0
Test             0
ODI              0
T20              0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 646 entries, 24 to 6806
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Name           646 non-null    object
 1   Date_Of_Birth  646 non-null    object
```

```
 2   Country         646 non-null    object
 3   Test            646 non-null    float64
 4   ODI             646 non-null    float64
 5   T20             646 non-null    float64
dtypes: float64(3), object(3)
memory usage: 35.3+ KB
```

```
# Convert Date_Of_Birth to datetime
df['Date_Of_Birth'] = pd.to_datetime(df['Date_Of_Birth'])
```

```
#summary of a DataFrame
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 646 entries, 24 to 6806
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Name           646 non-null    object
 1   Date_Of_Birth  646 non-null    datetime64[ns]
 2   Country        646 non-null    object
 3   Test           646 non-null    float64
 4   ODI            646 non-null    float64
 5   T20            646 non-null    float64
dtypes: datetime64[ns](1), float64(3), object(2)
memory usage: 35.3+ KB
```

inside query

```
df.isnull().sum()
```

```
Name              0
Date_Of_Birth     0
Country           0
Test              0
ODI               0
T20               0
dtype: int64
```

```
#total number of matches played by each player
df[['Test', 'ODI', 'T20']].sum(axis=1)
```

```
24       60.0
43      343.0
58       34.0
64        8.0
65      200.0
        ...
6739      6.0
6782     28.0
6791     20.0
```

```
6804     27.0
6806      8.0
Length: 646, dtype: float64
```

```
#Players who have playedmore than50 Testmatches
df[df['Test'] > 50]
```

	Name	Date_Of_Birth	Country	Test	ODI	T20
229	M M Ali	1987-06-18	England	68.0	138.0	82.0
321	H M Amla	1983-03-31	South Africa	124.0	181.0	44.0
336	J M Anderson	1982-07-30	England	183.0	194.0	19.0
448	Asad Shafiq	1986-01-28	Pakistan	77.0	60.0	10.0
463	R Ashwin	1986-09-17	India	95.0	116.0	65.0
...	...	...	...	...	...	...
6488	B J Watling	1985-07-09	New Zealand	75.0	28.0	5.0
6492	S R Watson	1981-06-17	Australia	59.0	190.0	58.0
6605	K S Williamson	1990-08-08	New Zealand	96.0	165.0	87.0
6669	U T Yadav	1987-10-25	India	57.0	75.0	9.0
6713	Younis Khan	1977-11-29	Pakistan	118.0	265.0	25.0

[126 rows x 6 columns]

```
#Find the player highest ODI score
df.loc[df['ODI'].idxmax(), 'Name']
```

```
'S R Tendulkar'
```

```
#Count the number of players from each country
df['Country'].value_counts()
```

```
Country
England         69
New Zealand     66
Sri Lanka       66
Pakistan        64
Australia       59
India           58
Zimbabwe        58
South Africa    57
West Indies     57
Bangladesh      54
Ireland         21
Afghanistan     17
Name: count, dtype: int64
```

```
#find players who were born before 1980
df[df['Date_Of_Birth'] < '1980-01-01']
```

	Name	Date_Of_Birth	Country	Test	ODI	T20
43	Abdul Razzaq	1979-12-02	Pakistan	46.0	265.0	32.0
64	Abdur Rauf	1978-12-09	Pakistan	3.0	4.0	1.0

```
99          A R Adams    1975-07-17  New Zealand    1.0   42.0   4.0
140       A B Agarkar    1977-12-04        India   26.0  191.0   4.0
187      Aizaz Cheema    1979-09-05     Pakistan    7.0   14.0   5.0
...               ...           ...          ...    ...    ...   ...
6322      D L Vettori    1979-01-27  New Zealand  113.0  295.0  34.0
6343        L Vincent    1978-11-11  New Zealand   23.0  102.0   9.0
6369        A C Voges    1979-10-04    Australia   20.0   31.0   7.0
6713      Younis Khan    1977-11-29     Pakistan  118.0  265.0  25.0
6804  Zulfiqar Babar     1978-12-10     Pakistan   15.0    5.0   7.0

[126 rows x 6 columns]
```

```
#count the numberof unique countries in dataset
df['Country'].nunique()
```

```
12
```

```
#checkfor playersborn in february
df[df['Date_Of_Birth'].dt.month == 2]
```

```
                 Name Date_Of_Birth      Country  Test    ODI   T20
727        H K Bennett    1987-02-22  New Zealand   1.0   19.0  11.0
886      M G Bracewell    1991-02-14  New Zealand   8.0   19.0  16.0
904          D M Bravo    1989-02-06  West Indies  56.0  122.0  26.0
914        T T Bresnan    1985-02-28      England  23.0   85.0  34.0
931          H C Brook    1999-02-22      England  12.0   15.0  29.0
...                ...           ...          ...   ...    ...   ...
6286  J D F Vandersay     1990-02-05    Sri Lanka   1.0   20.0  14.0
6339    R Vinay Kumar     1984-02-12        India   1.0   31.0   9.0
6363  K D K Vithanage     1991-02-26    Sri Lanka  10.0    6.0   3.0
6364        B V Vitori    1990-02-22     Zimbabwe   4.0   24.0  11.0
6613        G C Wilson    1986-02-05      Ireland   2.0  105.0  81.0

[66 rows x 6 columns]
```

```
#find if there are any players from India
df['Country'].str.contains('India').any()
```

```
True
```

```
#find the highest T20 Score
df.loc[df['T20'].idxmax(),'Name']
```

```
'R G Sharma'
```

```
#calculate the average Test score
df['Test'].mean()
```

```
27.9984520123839
```

```
#calculate the median ODI score
df['ODI'].median()
```

```
49.0
```

```
#find players from specific country, e.g, Pakistan
df[df['Country'] == 'Pakistan']
```

	Name	Date_Of_Birth	Country	Test	ODI	T20
43	Abdul Razzaq	1979-12-02	Pakistan	46.0	265.0	32.0
58	Abdullah Shafique	1999-11-20	Pakistan	16.0	12.0	6.0
64	Abdur Rauf	1978-12-09	Pakistan	3.0	4.0	1.0
66	Abdur Rehman	1980-03-01	Pakistan	22.0	31.0	8.0
170	Ahmed Shehzad	1991-11-23	Pakistan	13.0	81.0	59.0
...	...	...	...	...	...	...
6690	Yasir Shah	1986-05-02	Pakistan	48.0	25.0	2.0
6713	Younis Khan	1977-11-29	Pakistan	118.0	265.0	25.0
6735	Zahid Mahmood	1988-03-20	Pakistan	2.0	4.0	1.0
6804	Zulfiqar Babar	1978-12-10	Pakistan	15.0	5.0	7.0
6806	Zulqarnain Haider	1986-04-03	Pakistan	1.0	4.0	3.0

```
[64 rows x 6 columns]
```

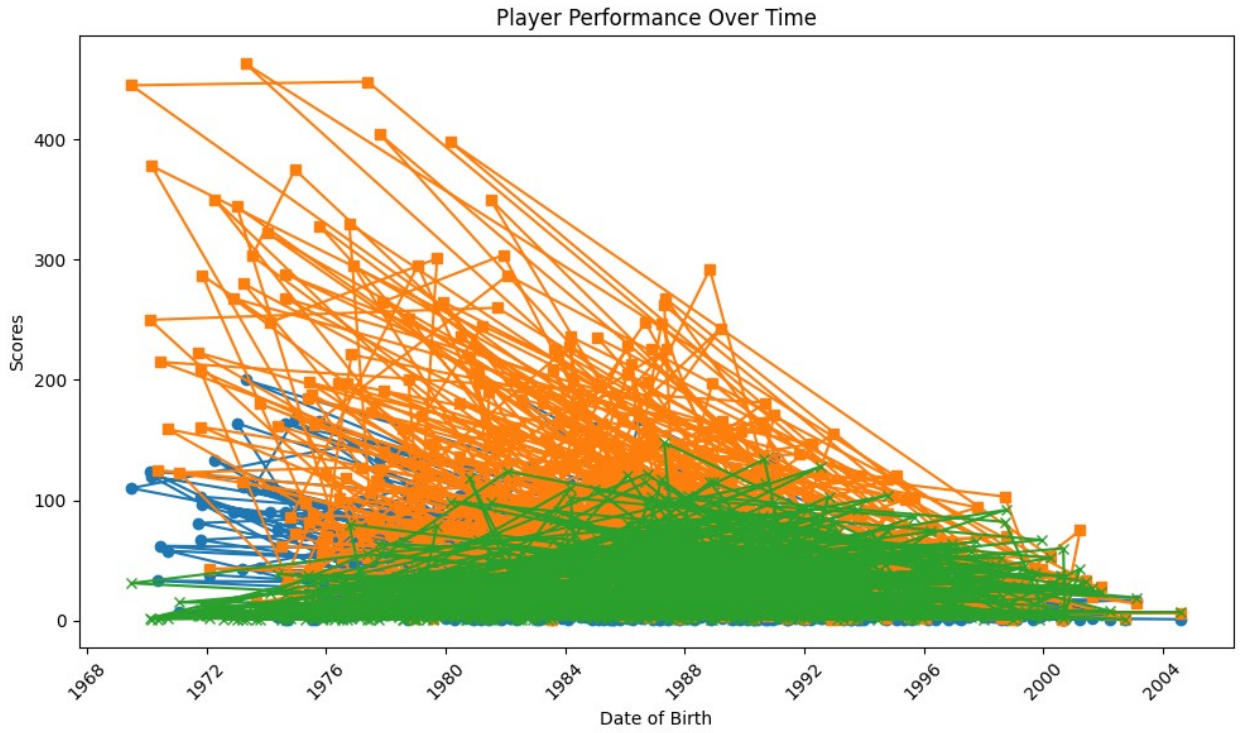
```
#Calculate the total number of players
df.shape[0]
```

```
646
```

matplotlib

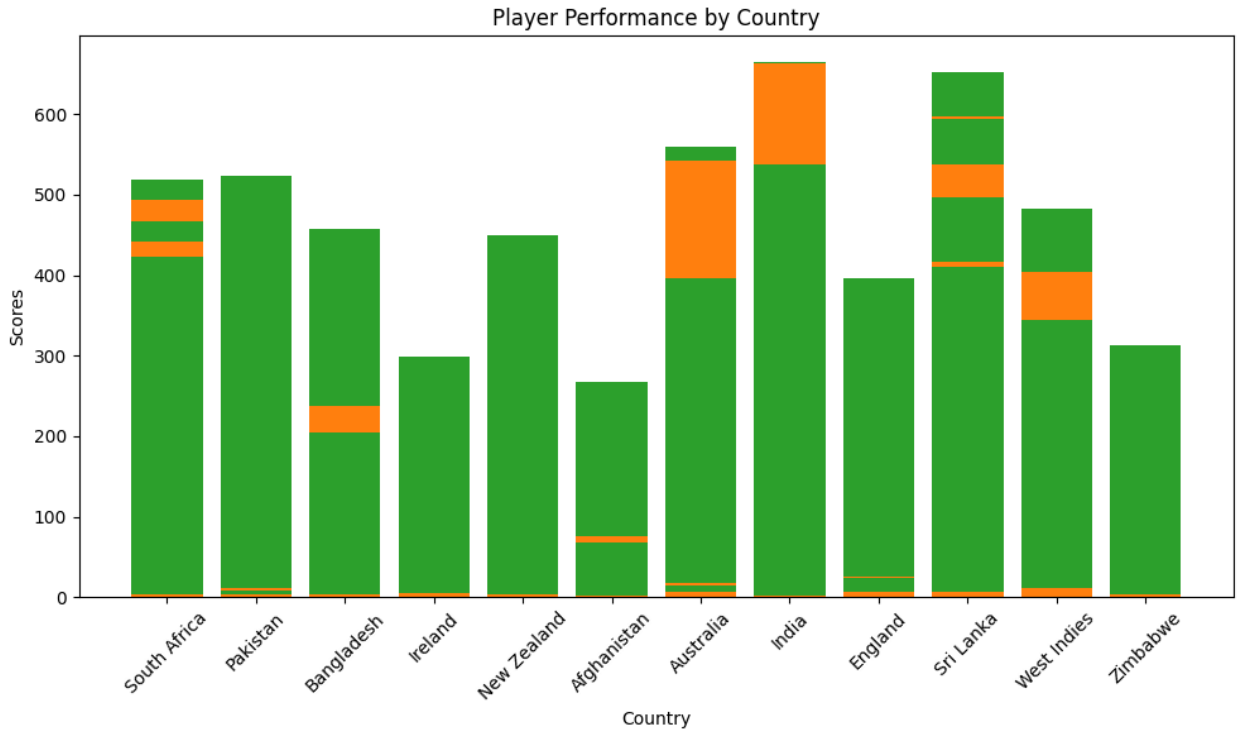
Lineplot

```
plt.figure(figsize=(10, 6))
plt.plot(df['Date_Of_Birth'], df['Test'], label='Test', marker='o')
plt.plot(df['Date_Of_Birth'], df['ODI'], label='ODI', marker='s')
plt.plot(df['Date_Of_Birth'], df['T20'], label='T20', marker='x')
plt.title('Player Performance Over Time')
plt.xlabel('Date of Birth')
plt.ylabel('Scores')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



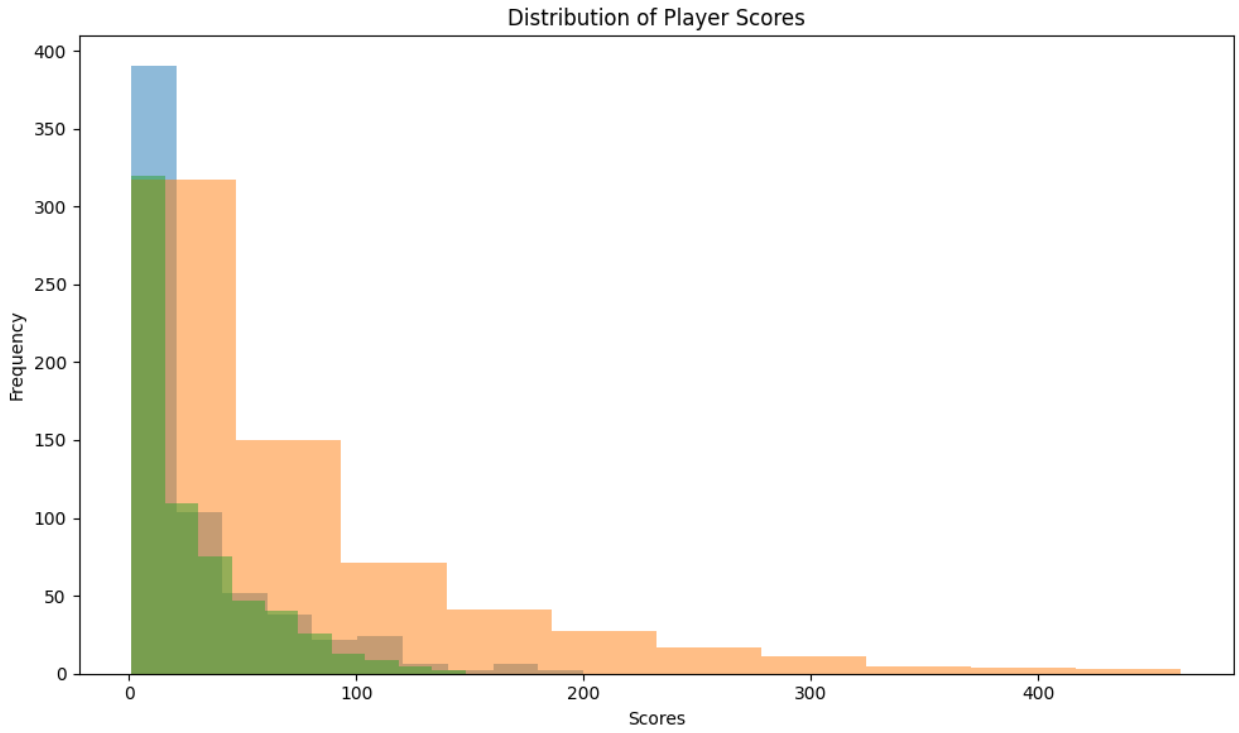
Barplot

```
plt.figure(figsize=(10, 6))
plt.bar(df['Country'], df['Test'], label='Test')
plt.bar(df['Country'], df['ODI'], bottom=df['Test'], label='ODI')
plt.bar(df['Country'], df['T20'], bottom=df['Test'] + df['ODI'],
label='T20')
plt.title('Player Performance by Country')
plt.xlabel('Country')
plt.ylabel('Scores')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



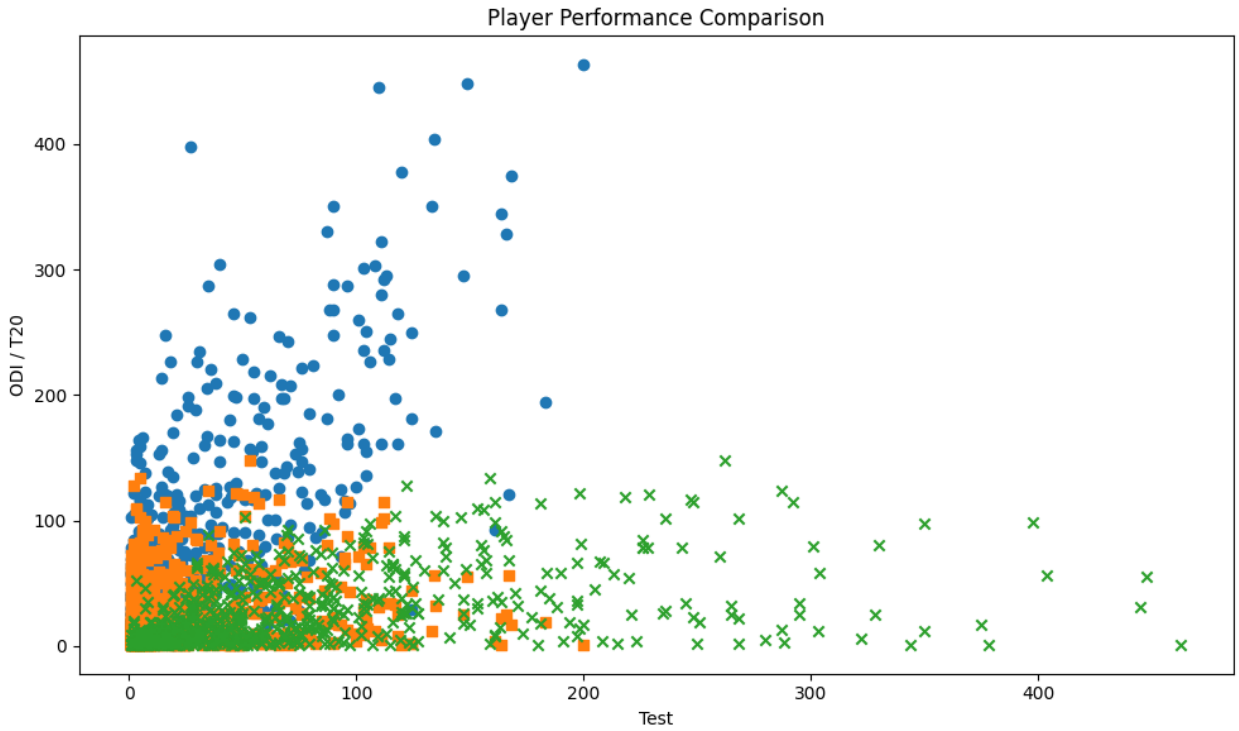
histogram

```
plt.figure(figsize=(10, 6))
plt.hist(df['Test'], bins=10, alpha=0.5, label='Test')
plt.hist(df['ODI'], bins=10, alpha=0.5, label='ODI')
plt.hist(df['T20'], bins=10, alpha=0.5, label='T20')
plt.title('Distribution of Player Scores')
plt.xlabel('Scores')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```



ScatterPlot

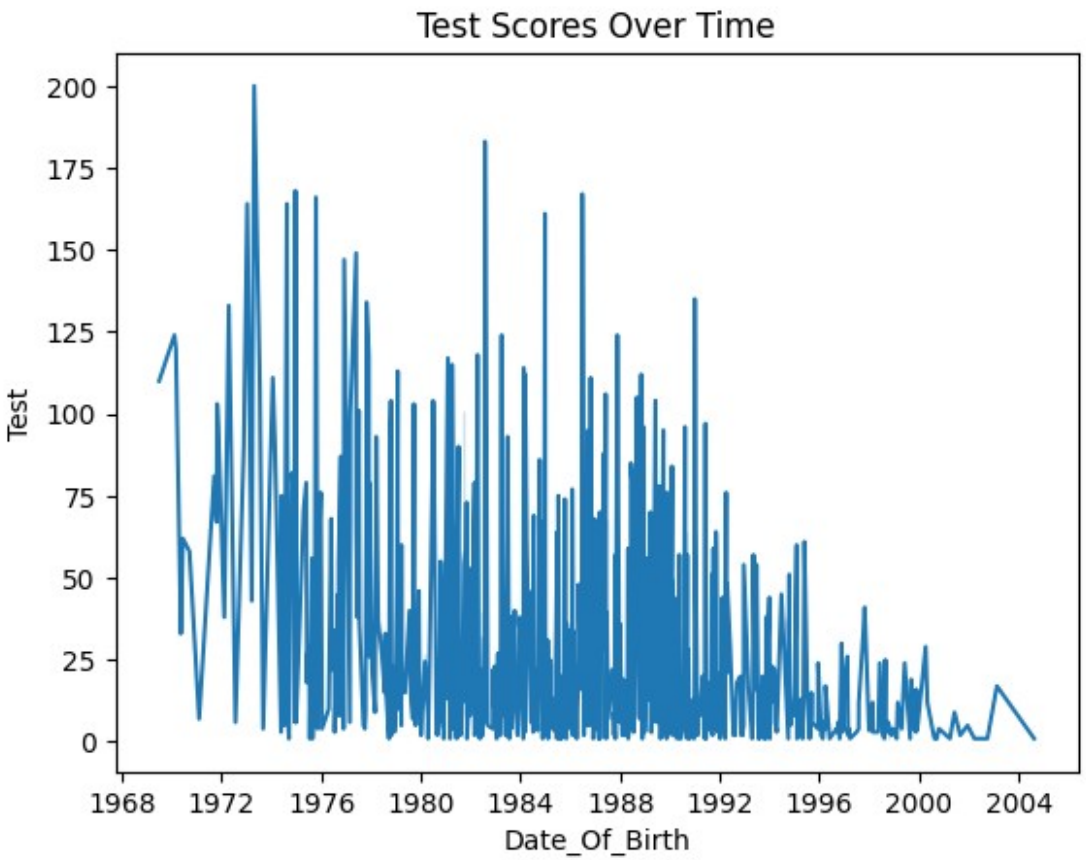
```
plt.figure(figsize=(10, 6))
plt.scatter(df['Test'], df['ODI'], label='Test vs ODI', marker='o')
plt.scatter(df['Test'], df['T20'], label='Test vs T20', marker='s')
plt.scatter(df['ODI'], df['T20'], label='ODI vs T20', marker='x')
plt.title('Player Performance Comparison')
plt.xlabel('Test')
plt.ylabel('ODI / T20')
plt.tight_layout()
plt.show()
```



seaborn

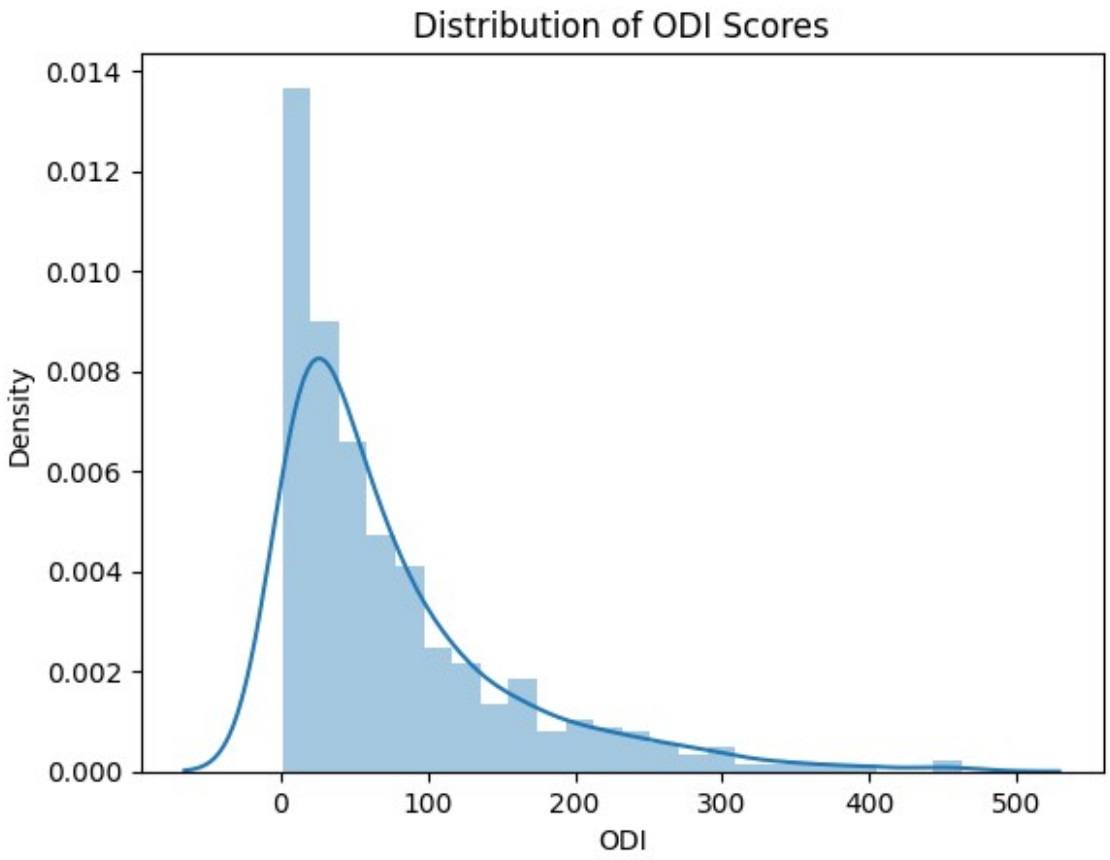
lineplot

```
sns.lineplot(x='Date_Of_Birth', y='Test', data=df)
plt.title('Test Scores Over Time')
plt.show()
```



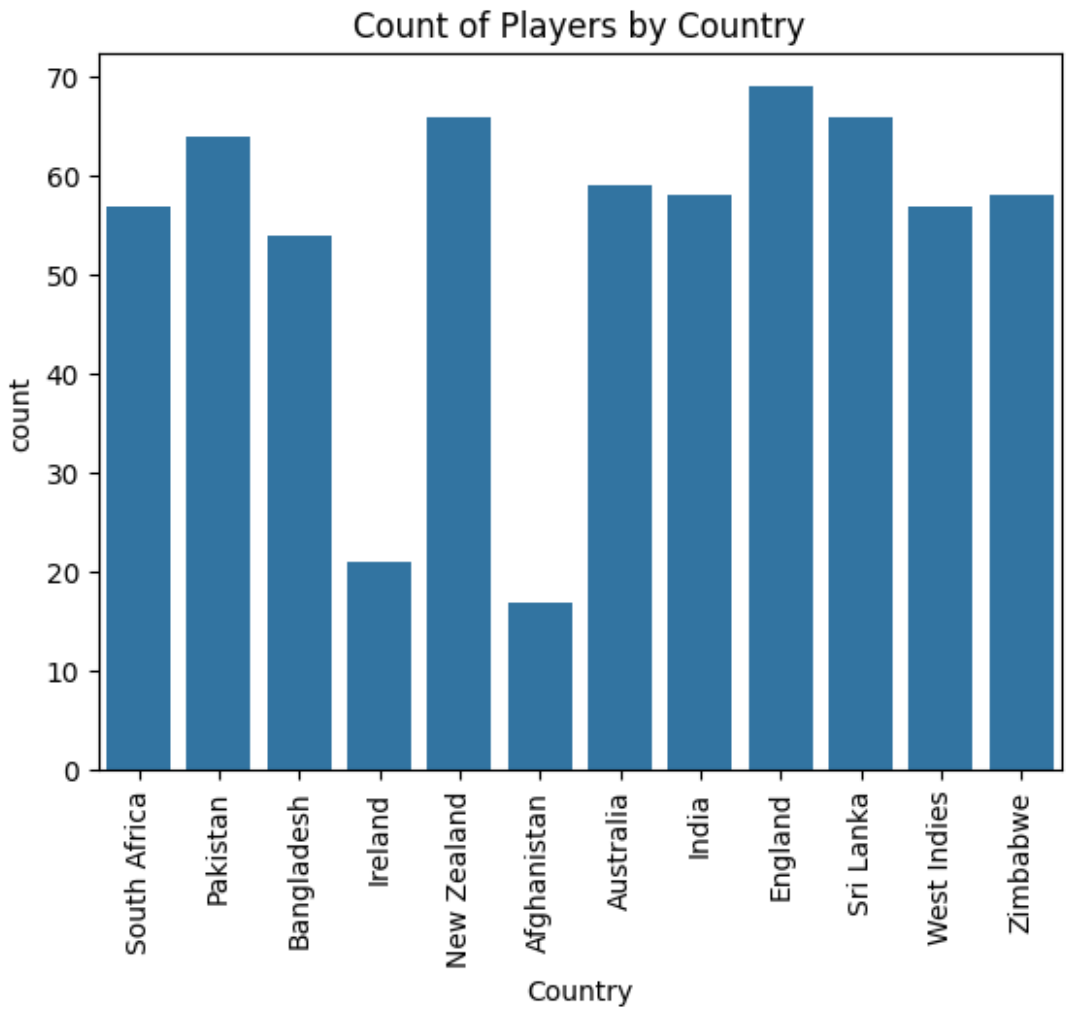
Distplot

```
sns.distplot(df['ODI'])
plt.title('Distribution of ODI Scores')
plt.show()
```



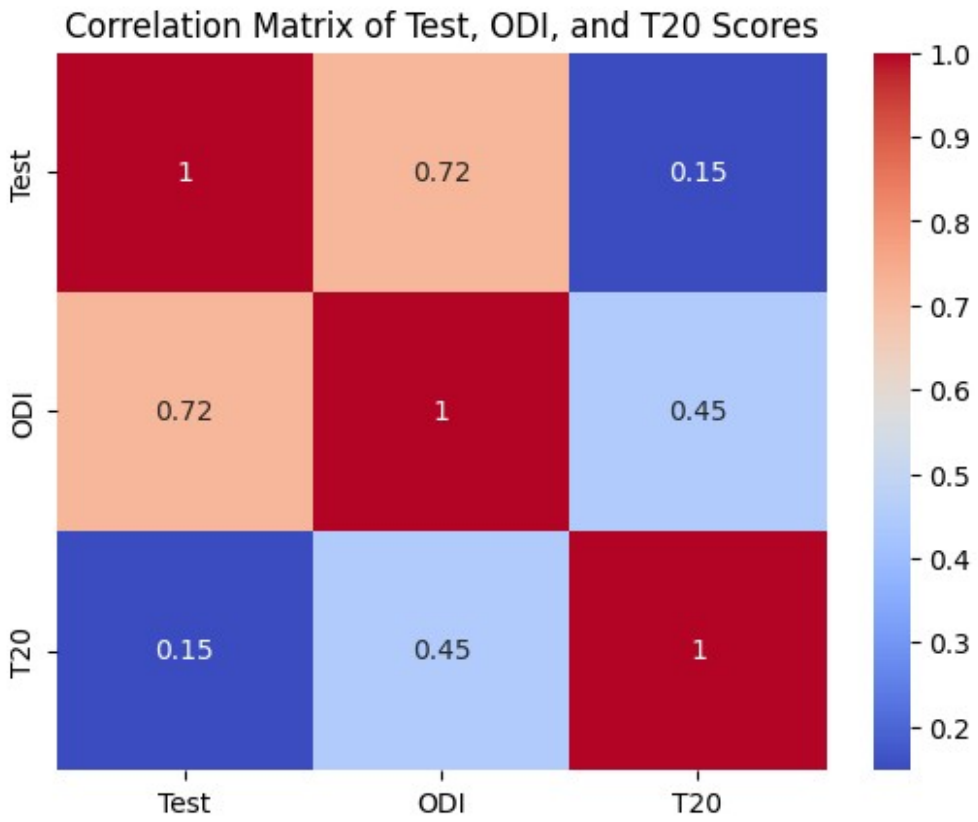
countplot

```
sns.countplot(x='Country', data=df)
plt.title('Count of Players by Country')
plt.xticks(rotation=90)
plt.show()
```



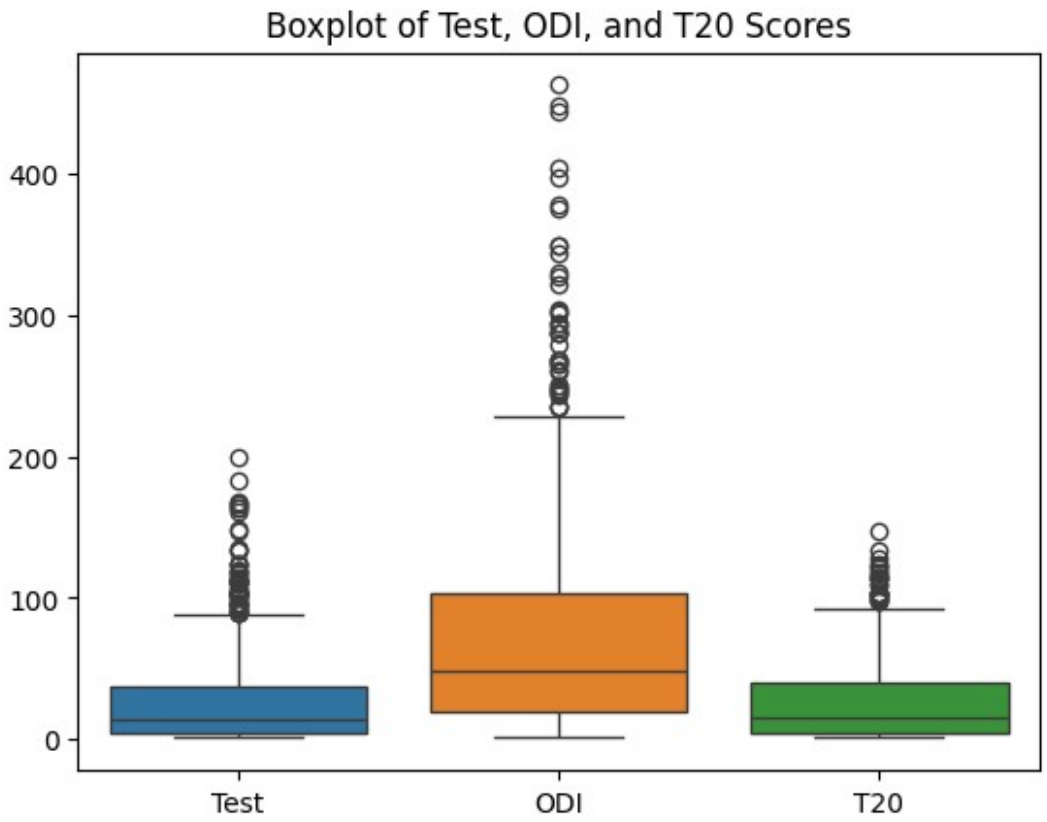
Heatmap

```
correlation_matrix = df[['Test', 'ODI', 'T20']].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of Test, ODI, and T20 Scores')
plt.show()
```



Boxplot

```
sns.boxplot(data=df[['Test', 'ODI', 'T20']])
plt.title('Boxplot of Test, ODI, and T20 Scores')
plt.show()
```



pairplot

```
sns.pairplot(df[['Test', 'ODI', 'T20']])
plt.show()
```

Parth_patil_24