CIS-9650 PROGRAMMING FOR ANALYTICS FINAL PROJECT

TEAM – 5

-Daniyal Mohammad

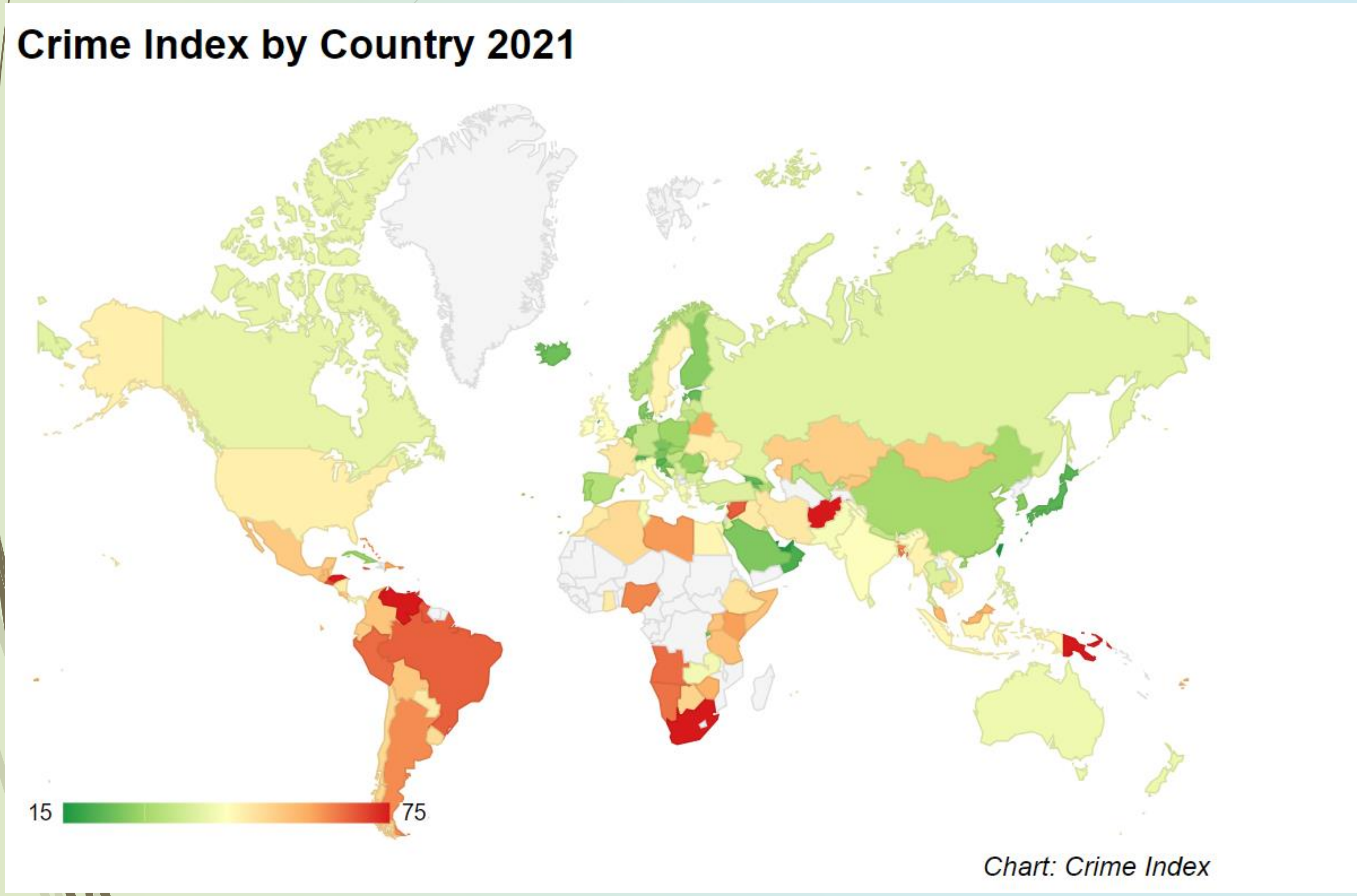
-Jay Sadrani

-Payal Surana

-Sadhvi Grover

Professor – Isaac Vaghefi

INTRODUCTION



- We're using a dataset from the **Global Initiative against Transnational Organized Crime** to analyze crime levels and resilience in 193 countries. Focusing on three key pillars: criminal markets, Criminal actors, and Resilience.

- Our goal is to help policymakers prioritize actions against crime and measure the effectiveness of their efforts.

- We're linking this data with GDP to understand how a country's economic performance relates to different types of crime.

IMPORTING DATA

- Utilizing Python's libraries including pandas, numpy, matplotlib, and seaborn, we have imported the necessary files to analyze and visualize our data effectively in our presentation
- We read and organized the Excel files, creating a structured data frame essential for our analysis in Jupyter Notebook.

In [54]:

```
#Importing Necessary Files
import pandas as pd
from pandas import Series, DataFrame
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [58]:

```
# Reading the Excel file into a DataFrame
gdp = pd.read_excel("gdp_per_capita.xlsx")
gdp
```

Out[58]:

	Country/Area	Year	Unit	GDP, Per Capita GDP - US Dollars
0	Afghanistan	2021	US$	372.548875
1	Albania	2021	US$	6396.461812
2	Algeria	2021	US$	3700.324058
3	Andorra	2021	US$	42066.041570
4	Angola	2021	US$	2044.218212
...	...	...	...	...
188	Venezuela (Bolivarian Republic of)	2021	US$	3965.034328
189	Viet Nam	2021	US$	3756.488901
190	Yemen	2021	US$	301.586433
191	Zambia	2021	US$	1094.501613
192	Zimbabwe	2021	US$	1507.994790

193 rows × 4 columns

In [59]:

```
# Reading the Excel file into a DataFrame, specifying the sheet name
crm = pd.read_excel("organized_crime.xlsx",sheet_name="2021_dataset")
crm
```

CLEANING UP THE DATA

In [30]:
```python
# Creating a set of countries from the "Country/Area" column in the GDP and Crime DataFrame
gdp_countries = set(gdp["Country/Area"])
crime_countries = set(crm["Country"])

gdp_countries_without_match = gdp_countries.difference(crime_countries)


crime_countries_without_match = crime_countries.difference(gdp_countries)
```

- During the data cleaning process, we identified countries in the GDP dataset that were missing from the crime dataset and vice versa. By addressing these inconsistencies, we ensured uniformity in country names across both datasets, facilitating accurate analysis.

```python
#Creating a new list to store incorrect spellings and their correct replacements for country names.
replacing_gdp_countries = []
for crime_countries in crime_countries_without_match:
  for gdp_countries in gdp_countries_without_match:
    if crime_countries in gdp_countries:
      tuple=(crime_countries,gdp_countries)
      replacing_gdp_countries.append(tuple)
replacing_gdp_countries
```

- To rectify inconsistencies between the two datasets, we created a list containing misspelled country names and their correct replacements. This allowed us to harmonize the country names across both datasets, ensuring accurate analysis.

In [32]:

```python
#in this step, we are adding missing countries names to a new list
# merging partially and not completely matching countries
#final list with missing and correct spellings
missing_gdp_countries = [
    ("Turkey", "Türkiye"),
    ("Korea, DPR", "Democratic People's Republic of Korea"),
    ("Vietnam", "Viet Nam"),
    ("Congo, Rep.", "Congo"),
    ("Korea, Rep.", "Republic of Korea"),
    ("Congo, Dem. Rep.", "Democratic Republic of the Congo"),
    ("St. Kitts and Nevis", "Saint Kitts and Nevis"),
    ("St. Vincent and the Grenadines", "Saint Vincent and the Grenadines"),
    ("St. Lucia", "Saint Lucia"),
    ("Czech Republic", "Czechia"),
    ("Laos", "Lao People's Democratic Republic")
]


replacement_gdp_countries = missing_gdp_countries + replacing_gdp_countries
```

- We compiled a list of missing country names and their correct spellings, merging them with the previously generated corrections for complete accuracy in our datasets.

In [10]:

```python
#renaming Country/Area to Country so that it matches with crime data and will help in merging the 2 datasets
crm.rename(columns={'Country': 'Country'}, inplace=True)
gdp.rename(columns={'Country/Area': 'Country'}, inplace=True)
gdp
```

Out[10]:

	Country	Year	Unit	GDP, Per Capita GDP - US Dollars
0	Afghanistan	2021	US$	372.548875
1	Albania	2021	US$	6396.461812
2	Algeria	2021	US$	3700.324058
3	Andorra	2021	US$	42066.041570
4	Angola	2021	US$	2044.218212
...	...	...	...	...
188	Venezuela (Bolivarian Republic of)	2021	US$	3965.034328
189	Viet Nam	2021	US$	3756.488901
190	Yemen	2021	US$	301.586433
191	Zambia	2021	US$	1094.501613
192	Zimbabwe	2021	US$	1507.994790

- We renamed the "Country/Area" column in the GDP dataset to simply "Country" to align it with the naming convention in the crime dataset, easing the merging process between the two datasets.

Run Markdown

In [11]:

```python
# replacing country names from GDP Dataframe with the ones from crime dataframe
for replacement in replacement_gdp_countries:
    new, old = replacement
    gdp.loc[gdp['Country'] == old, 'Country'] = new
gdp
```

Out[11]:

	Country	Year	Unit	GDP, Per Capita GDP - US Dollars
0	Afghanistan	2021	US$	372.548875
1	Albania	2021	US$	6396.461812
2	Algeria	2021	US$	3700.324058
3	Andorra	2021	US$	42066.041570
4	Angola	2021	US$	2044.218212
...	...	...	...	...
188	Venezuela	2021	US$	3965.034328
189	Vietnam	2021	US$	3756.488901
190	Yemen	2021	US$	301.586433
191	Zambia	2021	US$	1094.501613
192	Zimbabwe	2021	US$	1507.994790

193 rows × 4 columns

In [12]:

```python
#rounding off the GDP to two decimals to make sense of numbers better
gdp["GDP, Per Capita GDP - US Dollars"] = gdp["GDP, Per Capita GDP - US Dollars"].round(2)
gdp
```

Out[12]:

	Country	Year	Unit	GDP, Per Capita GDP - US Dollars
0	Afghanistan	2021	US$	372.55
1	Albania	2021	US$	6396.46
2	Algeria	2021	US$	3700.32
3	Andorra	2021	US$	42066.04
4	Angola	2021	US$	2044.22
...	...	...	...	...
188	Venezuela	2021	US$	3965.03
189	Vietnam	2021	US$	3756.49
190	Yemen	2021	US$	301.59
191	Zambia	2021	US$	1094.50
192	Zimbabwe	2021	US$	1507.99

193 rows × 4 columns

- After renaming, we replaced the old column of Country in the GDP dataset with a new column so that it is updated.
- we rounded off the GDP values to two decimal places for better clarity and ease of interpretation of the numbers.

MERGING TWO DATASETS

```
In [13]:
#Now, we are merging the two datasets after cleaning to create a accurate dataset.
merged_data = pd.merge(crm, gdp, left_on='Country', right_on='Country', how='left')
merged_data
```

Out[13]:

Criminal markets	Human trafficking	Human smuggling	Arms trafficking	Flora crimes	Fauna crimes	...	Law enforcement	Territorial integrity	Anti-money laundering	Economic regulatory capacity	Victim and witness support	Prevention	Non-state actors	Year	Unit	GDP, Per Capita GDP - US Dollars
6.40	7.0	9.0	9.0	4.0	3.0	...	3.0	6.5	2.0	4.0	4.0	3.5	3.5	2021	US$	9661.23
3.70	4.5	2.0	2.5	2.5	4.0	...	6.0	7.5	5.0	6.0	3.5	6.0	7.0	2021	US$	3293.23
6.00	4.5	4.0	8.0	3.5	7.5	...	5.0	6.0	6.0	6.0	4.0	5.0	7.0	2021	US$	7055.06
7.20	7.5	7.0	8.0	6.0	7.0	...	6.0	4.5	6.0	5.0	3.5	4.5	6.5	2021	US$	6104.14
6.20	7.0	6.0	5.5	7.0	5.5	...	5.0	5.0	4.5	4.0	3.5	3.0	4.5	2021	US$	6621.65
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4.20	5.0	3.0	8.5	1.5	2.0	...	6.0	6.5	4.5	4.5	5.0	6.0	6.5	2021	US$	5183.58
2.85	2.5	1.0	4.0	1.0	2.0	...	4.5	6.0	5.5	4.0	3.0	6.0	6.0	2021	US$	8440.03
3.95	4.0	5.5	3.5	2.0	3.0	...	6.0	7.5	6.5	5.0	6.0	6.0	6.0	2021	US$	29134.80
3.05	4.0	1.5	5.0	2.0	1.0	...	4.5	5.5	4.0	5.5	5.0	6.0	4.0	2021	US$	9824.06

```
In [14]:
#changing the name of dataset as now the order of countries is Alphabetical
organized_data = merged_data
organized_data = organized_data.sort_values(by=['Country'],ascending=[True])
```

- Now, we merge the two cleaned datasets to create an accurate dataset for analysis.
- Alternatively, to ensure the dataset is ordered alphabetically by country names, we labeled it as "organized_data" following the sorting process.

STATISTICAL ANALYSIS OF THE DATASET

- We used the describe function in our Organized dataset to provide an overview of each column, such as the total count, the mean value, the standard deviation, the minimum value, and the maximum value. For example, using describe, we learned that the mean for Criminal markets in the 2021 period was around 4.87, with a standard deviation of 1.32.

In [15]:

```
#Statistical Analysis of the recent dataset
organized_data.describe()
```

Out[15]:

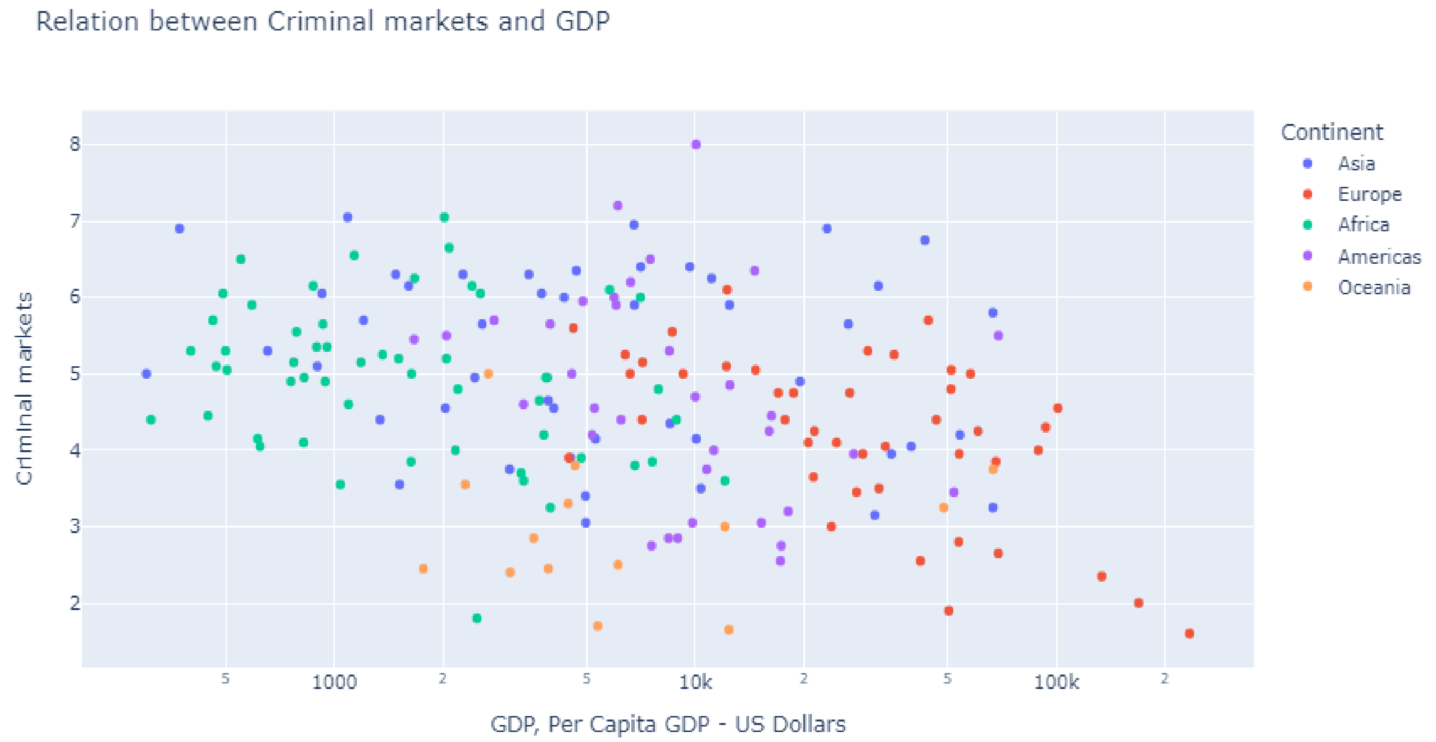
	Criminality	Criminal markets	Human trafficking	Human smuggling	Arms trafficking	Flora crimes	Fauna crimes	Non-renewable resource crimes	Heroin trade	Cocaine trade	...	Judicial system and detention	Law enforcement	
count	193.000000	193.000000	193.000000	193.00000	193.000000	193.000000	193.000000	193.000000	193.000000	193.000000	...	193.000000	193.000000	1
mean	4.872383	4.650777	5.582902	4.76943	4.919689	3.878238	4.634715	4.505181	3.974093	4.523316	...	4.593264	4.911917	
std	1.326322	1.272582	1.679648	1.91416	2.105307	2.315469	1.921639	2.432950	2.060757	2.016398	...	1.831895	1.768507	
min	1.540000	1.600000	1.500000	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.500000	1.500000	
25%	4.000000	3.850000	4.500000	3.00000	3.000000	2.000000	3.500000	2.000000	2.000000	3.000000	...	3.000000	4.000000	
50%	4.900000	4.750000	5.500000	5.00000	5.000000	3.500000	4.500000	4.000000	4.000000	4.500000	...	4.500000	5.000000	
75%	5.890000	5.650000	7.000000	6.50000	6.500000	6.000000	6.000000	6.500000	5.500000	6.000000	...	6.000000	6.000000	
max	7.750000	8.000000	9.500000	9.50000	9.500000	8.500000	9.000000	9.500000	9.500000	9.500000	...	9.000000	9.000000	

8 rows × 32 columns

DATA VISUALIZATION

In [39]:

```python
#importing plotly.express to understand data through different charts
import plotly.express as px
# scatter chart of criminal markets and GDP
fig = px.scatter(organized_data,
                 y="Criminal markets",
                 x='GDP, Per Capita GDP - US Dollars',
                 log_x=True,
                 color="Continent",
                 hover_name="Country",
                 title="Relation between Criminal markets and GDP")

fig.show()
```

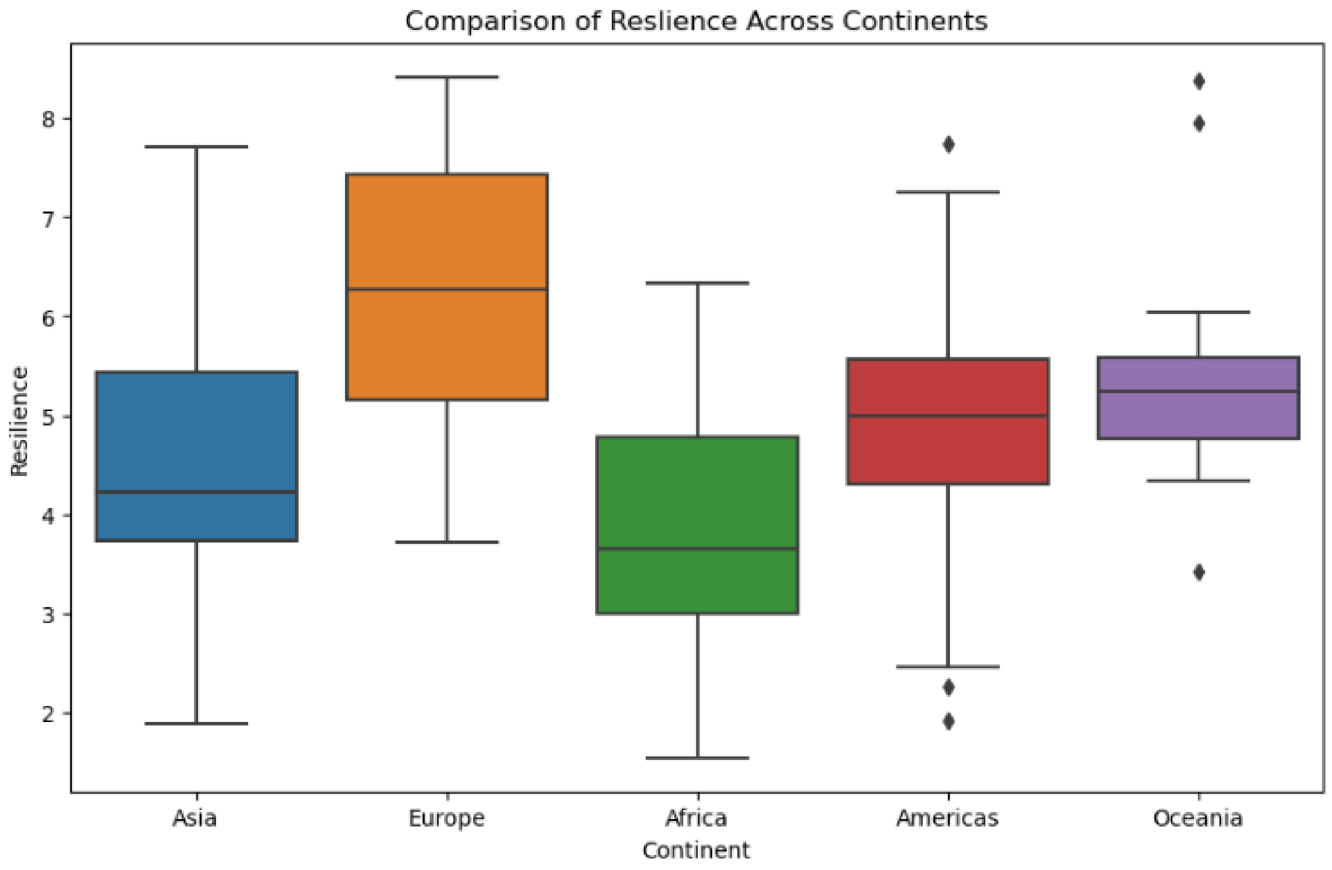


- **We explored scatterplot in Plotly Express** and visualized the relationship between criminal market ratings and GDP per capita, with hover-over functionality enabling detailed exploration of country-specific data.

- The graph depicts varying criminal market ratings across countries irrespective of GDP per capita, indicating a weak correlation. For instance, Afghanistan exhibits a high rating despite its low GDP per capita compared to the UAE.

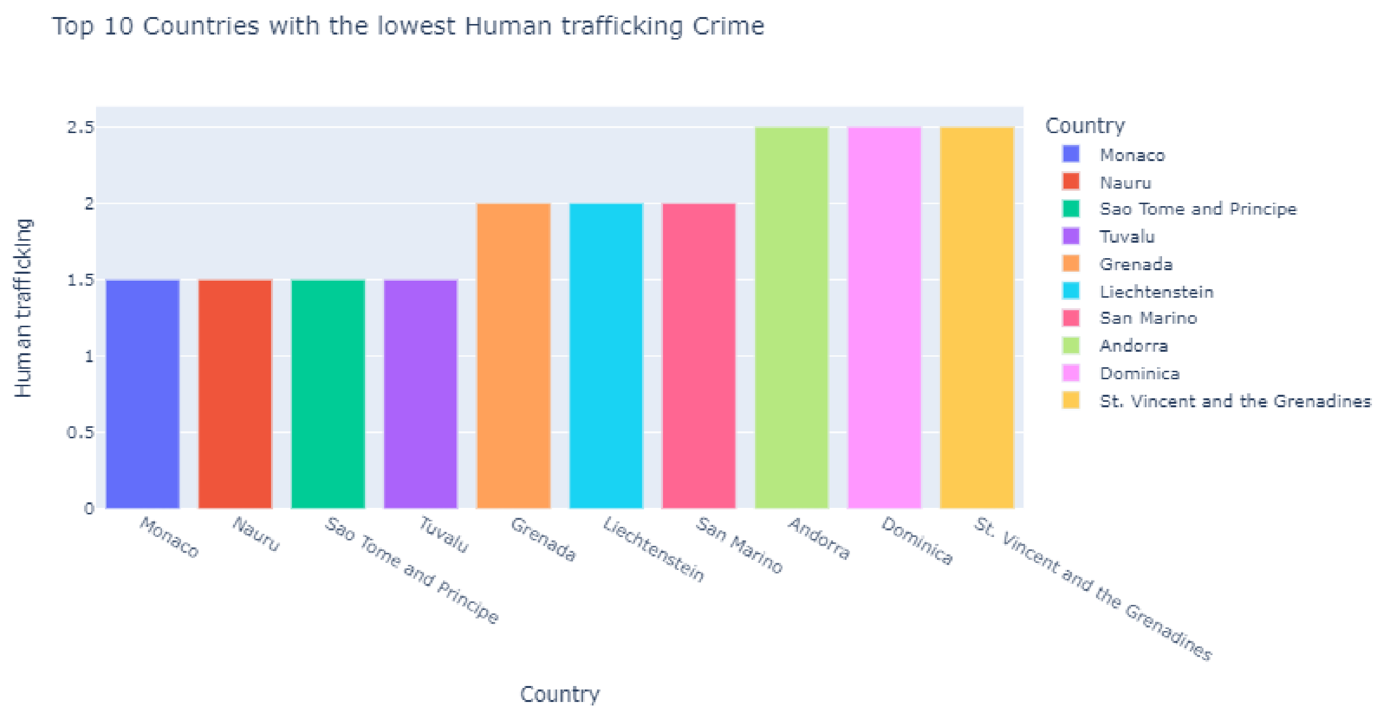
In [75]:

```python
# To define a outlier function we used Boxplot and chose Resilience with Continents
plt.figure(figsize=(10, 6))

sns.boxplot(x='Continent', y='Resilience', data=organized_data)
plt.title('Comparison of Reslience Across Continents')
plt.xlabel('Continent')
plt.ylabel('Resilience')

plt.show()
```



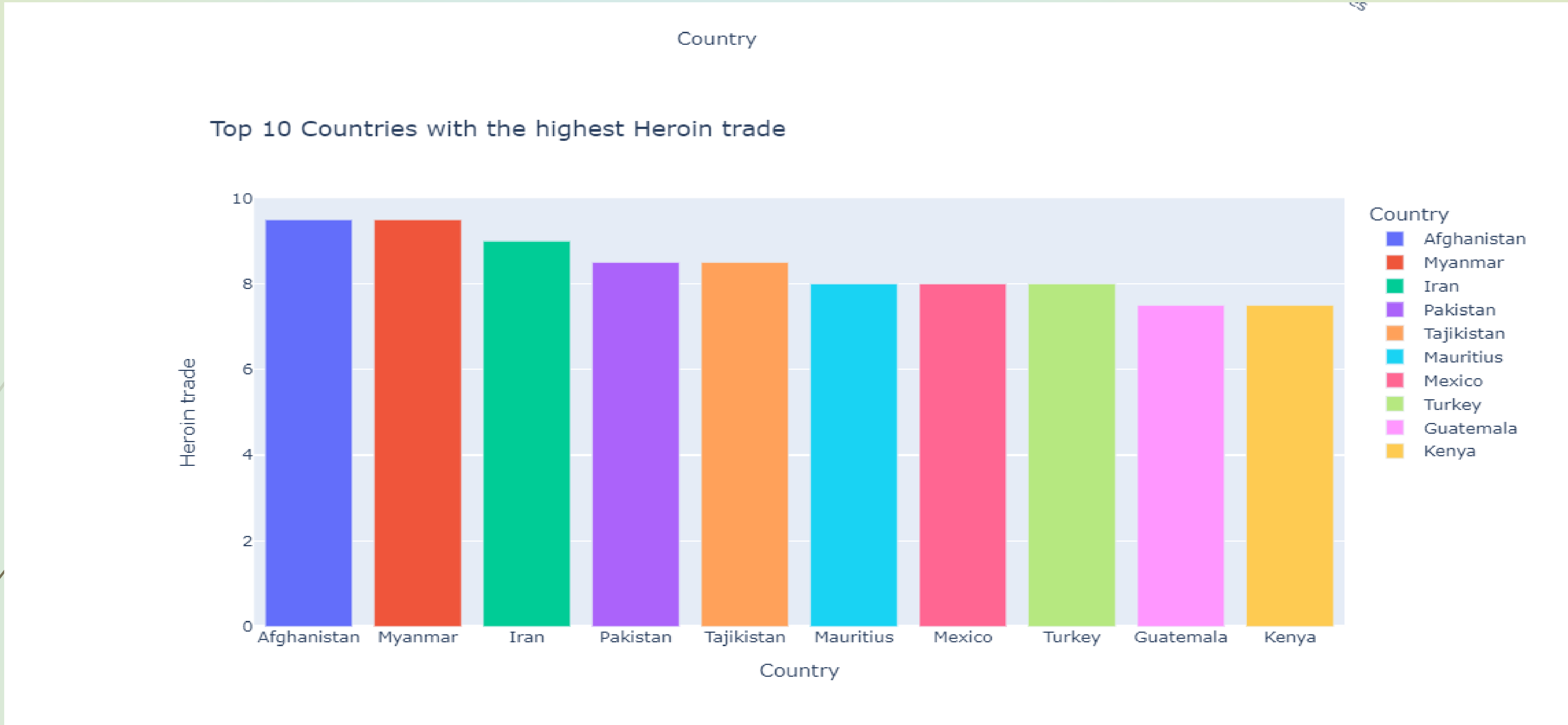
- We understood the outlier topic by using a **boxplot**

```
In [17]:
##crime_insights: Explore the top 10 countries with the lowest human trafficking crime and the highest heroin trade, offerin
crime = "Heroin trade"
crime_= "Human trafficking"

lowest_crime = organized_data.nsmallest(10, crime_)
highest_crime = organized_data.nlargest(10, crime)

px.bar(lowest_crime, x='Country', y=crime_, title="Top 10 Countries with the lowest Human trafficking Crime", color='Country


px.bar(highest_crime, x='Country', y=crime, title="Top 10 Countries with the highest Heroin trade", color='Country').show()
```



- **Through Plotly Express, we used a Bar chart** to present the top 10 countries with the lowest human trafficking crime and the highest heroin trade, offering valuable insights into global crime trends.

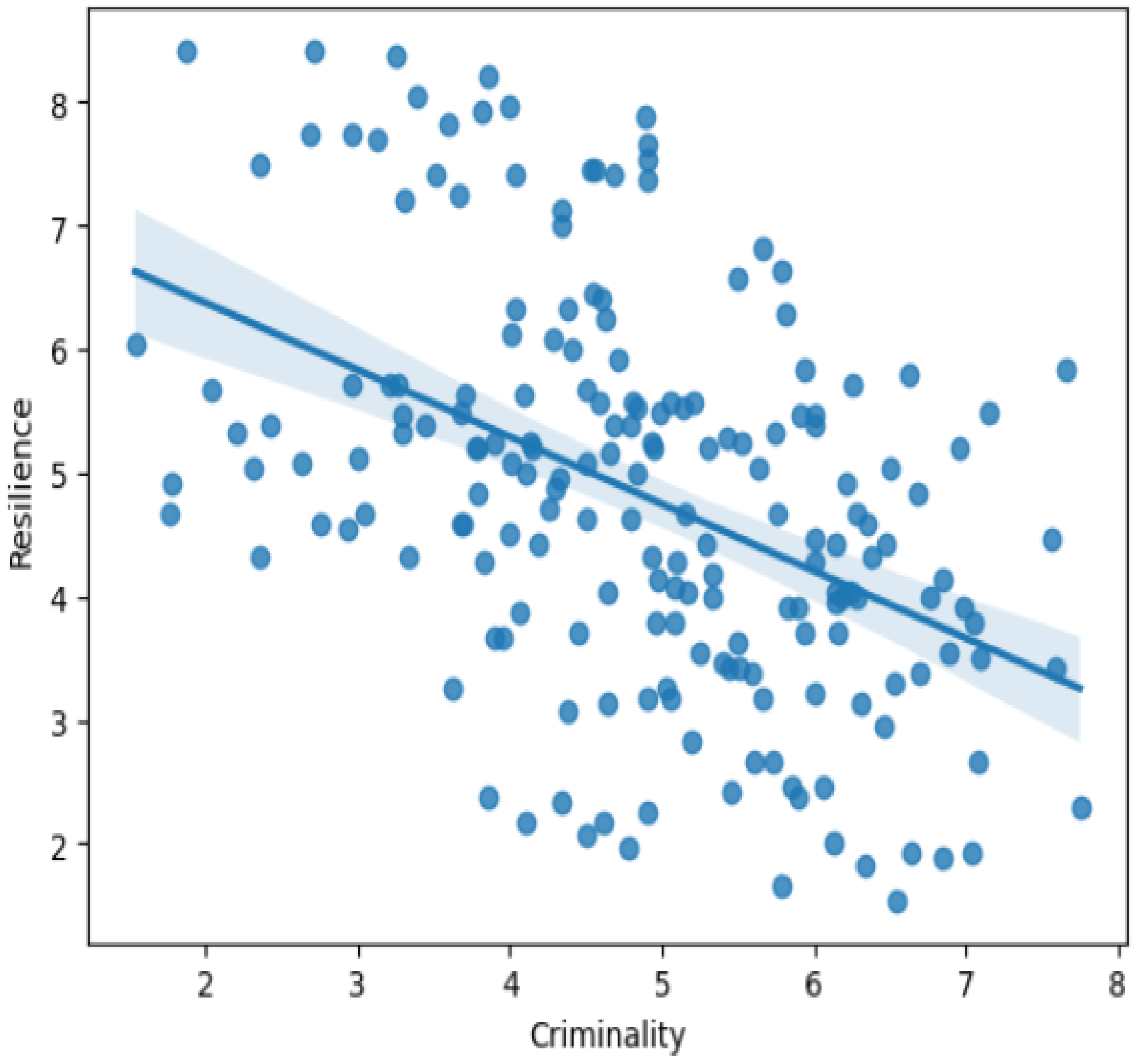
- We can see here that Monaco leads with the lowest rate at 1.5, suggesting effective measures in combating human trafficking. Nauru and Tuvalu follow closely, indicating relatively safer environments in these nations.



- This graph suggests that Afghanistan has the highest rate of heroin trade, indicating considerable difficulties in addressing this issue. Additionally, Myanmar, Iran, and Pakistan also have high rates, highlighting the extensive presence of this illegal trade in these countries.

```
In [18]:
# Visualizing the relationship between criminality and resilience, p
sns.regplot(x = 'Criminality',y='Resilience',data = organized_data)
```

```
Out[18]: <Axes: xlabel='Criminality', ylabel='Resilience'>
```

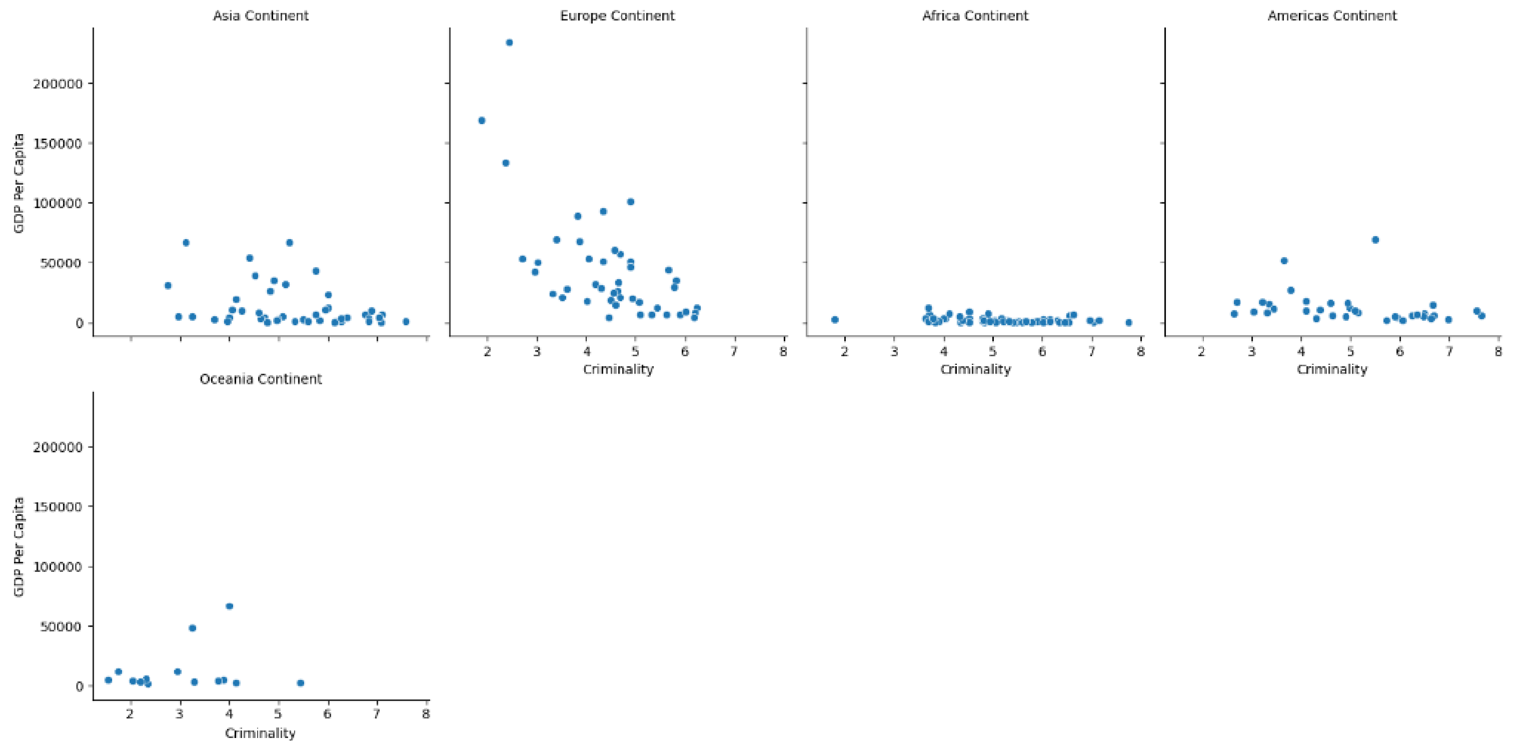


- We used **Regplot in seaborn** to understand the relationship between Criminality and Resilience.
- we understand a potential negative relationship between Criminality and Resilience, indicating that regions with higher criminality might exhibit lower resilience.
- However, the scattered distribution of data points implies that resilience scores can be influenced by various factors beyond just criminality.

In [19]:

```
##visualizing using sns.FacetGrid to create a scatterplot grid, visualizing the relationship between criminality and GDP per
g = sns.FacetGrid(organized_data, col='Continent', col_wrap=4, height=4)
g.map(sns.scatterplot, 'Criminality', 'GDP, Per Capita GDP - US Dollars')

g.set_titles("{col_name} Continent")
g.set_axis_labels('Criminality', 'GDP Per Capita')
g.add_legend()
plt.show()

```

```
C:\Users\Data\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning:

The figure layout has changed to tight
```



- Here, we used **Seaborn's FacetGrid** to generate a scatterplot grid, examining the association between criminality and GDP per capita across various continents.

- In Africa, Criminality varies despite the same GDP except for an outlier. Whereas, Europe shows that the higher the GDP, the lower the rate of criminality.

- In contrast to the Oceania Continent, the rest of the continents i.e Asia and Americas have a higher rate of criminality and low GDP.

REGRESSION ANALYSIS

- We imported **statsmodels.api to get linear regression model to** analyze the relationship between GDP per capita with Criminal Markets, Criminal actors, and Resilience.
- The regression model explains about 35.9% of the variance in the dependent variable, GDP Per Capita, as indicated by the R-squared value.
- The model demonstrates statistical significance with a low F-statistic p-value of 3.48e-18 .
- While Resilience significantly influences GDP per capita (p = 0), the impact of other factors like Criminal Markets, and Criminal Actors remains less significant.

REGRESSION ANALYSIS

In [24]:

```
#Analyzing the impact of various dimensions of organized crime on GDP per capita using multiple linear regression.
# We defined independent variables
# We then defined the dependent variable
# Then fit the multiple linear regression model
import statsmodels.api as sm

X = sm.add_constant(organized_data[['Criminal markets', 'Criminal actors', 'Resilience']])

y = organized_data['GDP, Per Capita GDP - US Dollars']

model = sm.OLS(y, X).fit()

print(model.summary())
```

```
                                  OLS Regression Results
============================================================================================
Dep. Variable:     GDP, Per Capita GDP - US Dollars   R-squared:                       0.359
Model:                                          OLS   Adj. R-squared:                  0.349
Method:                               Least Squares   F-statistic:                     35.36
Date:                              Sun, 12 May 2024   Prob (F-statistic):           3.48e-18
Time:                                      14:11:10   Log-Likelihood:                -2212.8
No. Observations:                               193   AIC:                             4434.
Df Residuals:                                   189   BIC:                             4447.
Df Model:                                         3
Covariance Type:                          nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const            -1.872e+04   1.09e+04     -1.720      0.087   -4.02e+04    2743.175
Criminal markets -4519.4033   2181.406     -2.072      0.040   -8822.433    -216.373
Criminal actors   1704.6631   1910.178      0.892      0.373   -2063.344    5472.670
Resilience        9965.9236   1190.558      8.371      0.000    7617.435    1.23e+04
==============================================================================
Omnibus:                      216.121   Durbin-Watson:                   2.107
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8493.186
Skew:                           4.372   Prob(JB):                         0.00
Kurtosis:                      34.300   Cond. No.                         56.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Daniyal Mohammad, Jay Sadrani, Payal Surana, Sadhvi Grover