STUDENT: TEJASHRI DENDI **UIN**: 670653291

Visualization Tool Report

World Statistics 2023 Economic Analysis

Visualization Tool Report

Seaborn is a library that allows you to create statistical analysis visualizations of data. It uses Matplotlib underneath the plot graphs. It provides beautiful default styles and color palettes. It offers a variety of powerful tools for visualizing data, including scatter plots, line plots, bar plots, heat maps, and many more. It also provides support for advanced statistical analysis, such as regression analysis, distribution plots, and categorical plots.

Installing and getting started

The basic invocation of pip will install seaborn and, if necessary, its mandatory dependencies. It is possible to include optional dependencies that give access to a few advanced features:

pip install seaborn[stats]

The library is also included as part of the Anaconda distribution, and it can be installed with conda.
Mandatory dependencies – NumPy, pandas, matplotlib

Optional dependencies -

1) statsmodels, for advanced regression plots

2) scipy, for clustering matrices and some advanced options

3) fastcluster, faster clustering of large matrices

Dataset Description: Wikipedia World Statistics (2023)

This dataset is obtained from kaggle website. It provides a comprehensive snapshot of global country statistics for the year 2023. It was scraped from various Wikipedia pages using Beautiful Soup, consolidating key indicators and metrics for 142 countries. The dataset covers diverse aspects such as land area, water area, Human Development Index (HDI), GDP forecasts, internet usage, and population changes for exploratory data analysis and research in fields such as economics, demographics, and international relations to gain insights into the socio-economic dynamics of countries worldwide.

Key Columns and Metrics:

Country: The name of the country.
Total in km2: Total area of the country.
Land in km2: Land area excluding water bodies.

Water in km2: Area covered by water bodies.
Water %: Percentage of the total area covered by water.
HDI: Human Development Index, a measure of a country's overall achievement in its social and economic dimensions.
%HDI Growth: Percentage growth in HDI.
IMF Forecast GDP(Nominal): International Monetary Fund's forecast for Gross Domestic Product in nominal terms.
World Bank Forecast GDP(Nominal): World Bank's forecast for Gross Domestic Product in nominal terms.
UN Forecast GDP(Nominal): United Nations' forecast for Gross Domestic Product in nominal terms.
IMF Forecast GDP(PPP): IMF's forecast for Gross Domestic Product in purchasing power parity terms.
World Bank Forecast GDP(PPP): World Bank's forecast for Gross Domestic Product in purchasing power parity terms.
CIA Forecast GDP(PPP): Central Intelligence Agency's forecast for Gross Domestic Product in purchasing power parity terms.
Internet Users: Number of internet users in the country.
UN Continental Region: Continental region classification by the United Nations.
UN Statistical Subregion: Statistical subregion classification by the United Nations.
Population 2022: Population of the country in the year 2022.
Population 2023: Population of the country in the year 2023.
Population %Change: Percentage change in population from 2022 to 2023.

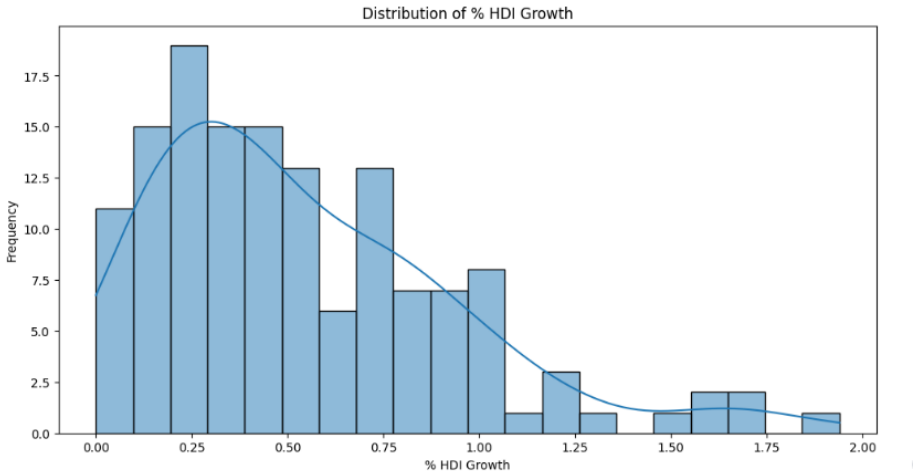
Variable selection:

'Total in km2', 'Land in km2', 'Internet Users', 'Population 2022', 'IMF Forecast GDP(Nominal)', 'HDI' are the variables selected from the dataset as these are playing vital role in the prediction of GDI and HDI.

Basic graphs:

Histplot of distribution of %HDI Growth:

```
# Plotting HDI Growth
plt.figure(figsize=(12, 6))
sns.histplot(df['%HDI Growth'], kde=True, bins=20)
plt.title('Distribution of % HDI Growth')
plt.xlabel('% HDI Growth')
plt.ylabel('Frequency')
plt.show()
```

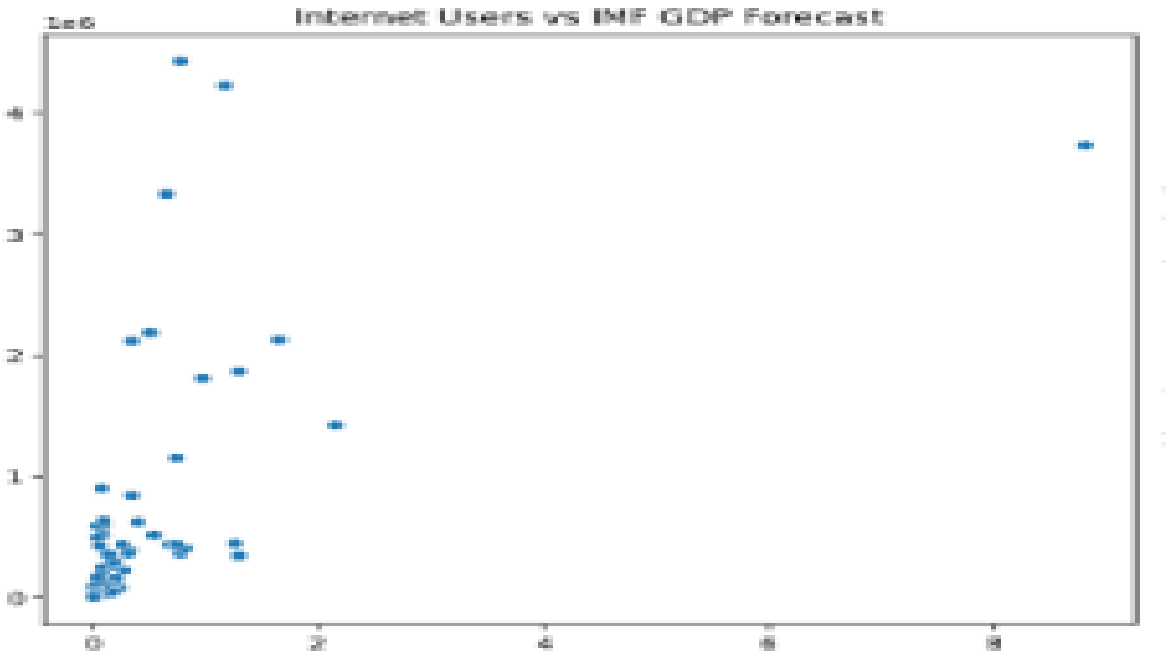


The above plot is right skewed.

<u>Scatterplot between Internet Users and IMF Forecast GDP (Nominal) :</u>

wikidata=df.dropna()

```
# GDP and Internet Penetration
# Scatter plot to investigate the relationship between Internet Users and GDP forecasts
fig, axes = plt.subplots(1, 3, figsize=(18, 6))

sns.scatterplot(data=wiki_data, x='Internet Users', y='IMF Forecast GDP(Nominal)', ax=axes
[0])
axes[0].set_title('Internet Users vs IMF GDP Forecast')
```

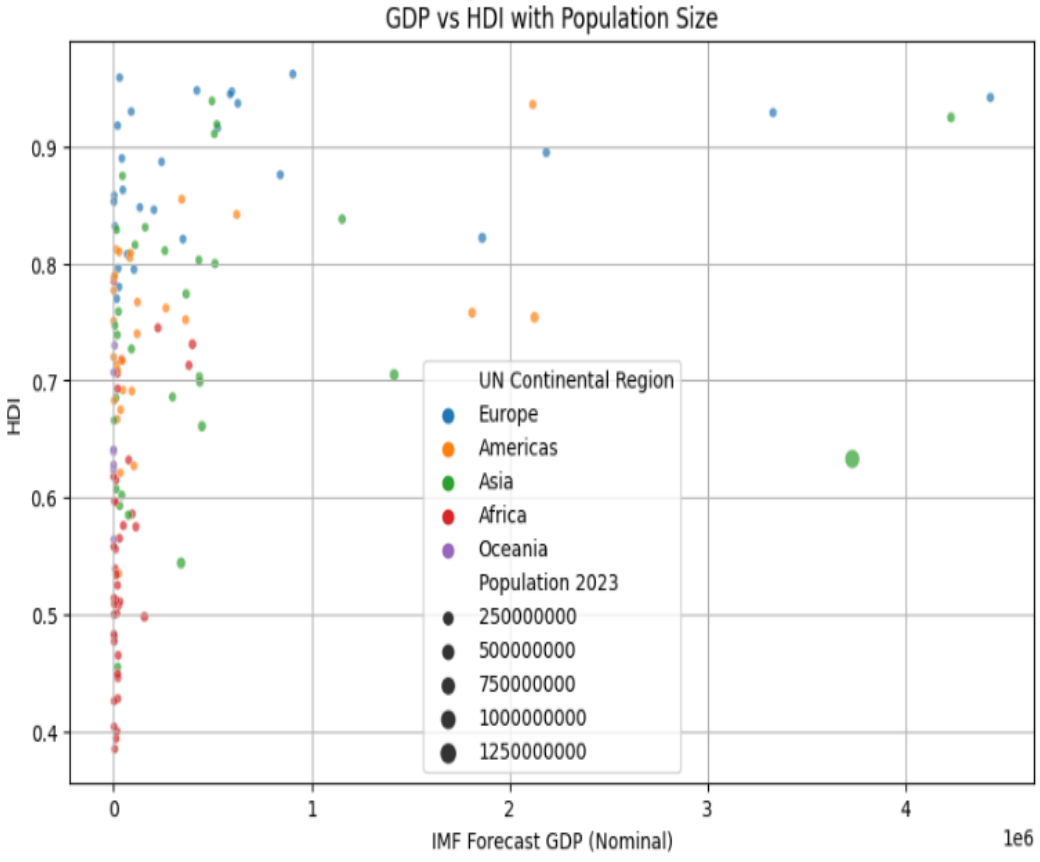


Advanced techniques:

a) Used legend (), grid (), xlabel (), ylabel(), title():

Scatterplot between IMF Forecast GDP(Nominal) and HDI:

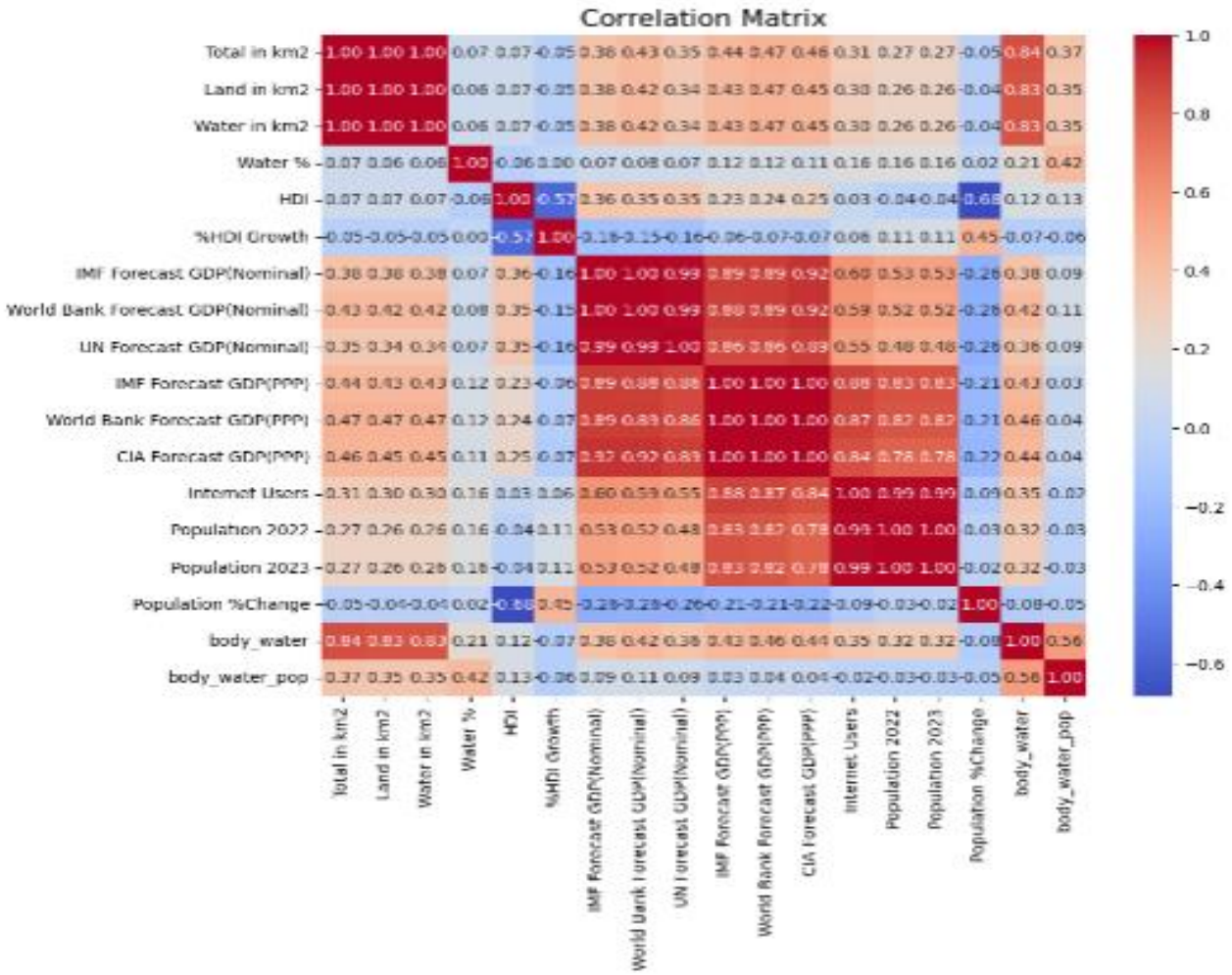
```
# Scatter Plot with GDP and HDI
plt.figure(figsize=(10, 6))
sns.scatterplot(x='IMF Forecast GDP(Nominal)', y='HDI', size='Population 2023',
                hue='UN Continental Region', data=df, alpha=0.7)
plt.title('GDP vs HDI with Population Size')
plt.xlabel('IMF Forecast GDP (Nominal)')
plt.ylabel('HDI')
plt.legend()
plt.grid(True)
plt.show()
```



b) Used set_title(), annot, fmt :

Correlation between numerical columns:

```
numerical_columns = df.select_dtypes(include='number')
correlation_matrix = numerical_columns.corr()

plt.figure(figsize=(10, 8))
heatmap = sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
heatmap.set_title('Correlation Matrix', fontsize=16)
plt.show()
```



From the above, it is clear that Internet users, population 2022, 2023 and Word Bank forecast (GDP, PPP), UN Forecast (GDP, PPP) are more correlated with IMF Forecast GDP (Nominal) when compared to other features. % HDI Growth is more correlated with HDI when compared to others.

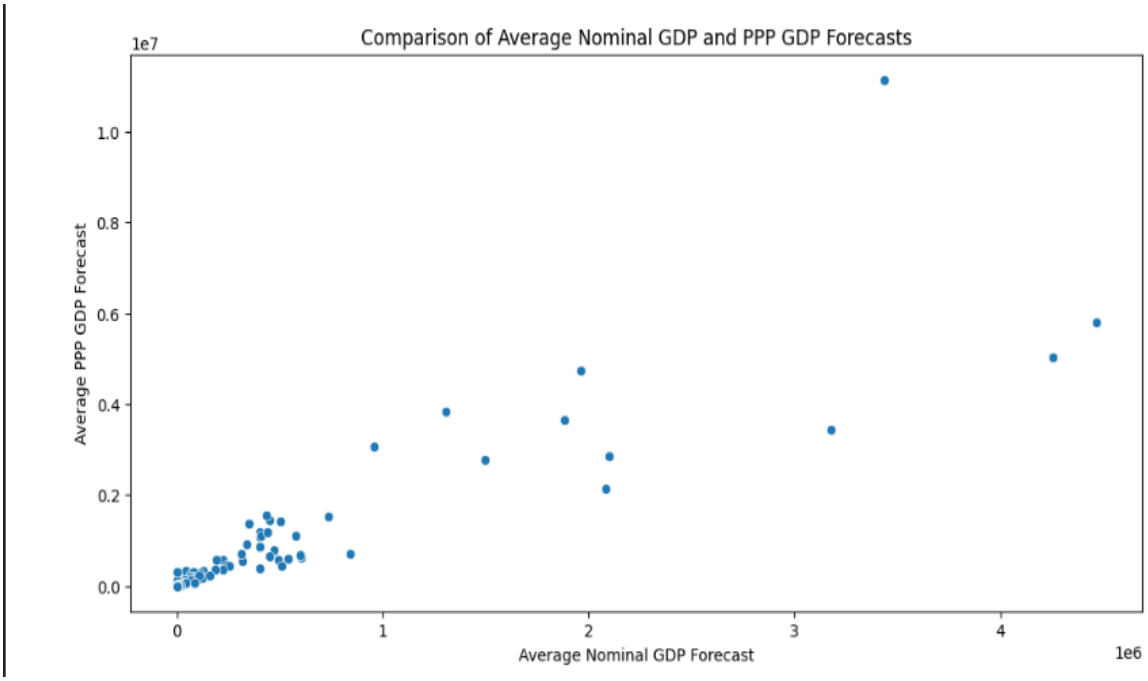
Observing the relationship between variables using advanced techniques like xscale(), yscale(), tight_layout():

```
# --- Comparative Analysis ---
# Compare different GDP forecasts for Nominal and PPP

# Average Nominal GDP forecast by different organizations
df['Average Nominal GDP Forecast'] = df[['IMF Forecast GDP(Nominal)',
                                         'World Bank Forecast GDP(Nominal)',
                                         'UN Forecast GDP(Nominal)']].mean(axis=1)
```

```
# Average PPP GDP forecast by different organizations
df['Average PPP GDP Forecast'] = df[['IMF Forecast GDP(PPP)',
                                     'World Bank Forecast GDP(PPP)',
                                     'CIA Forecast GDP(PPP)']].mean(axis=1)
```

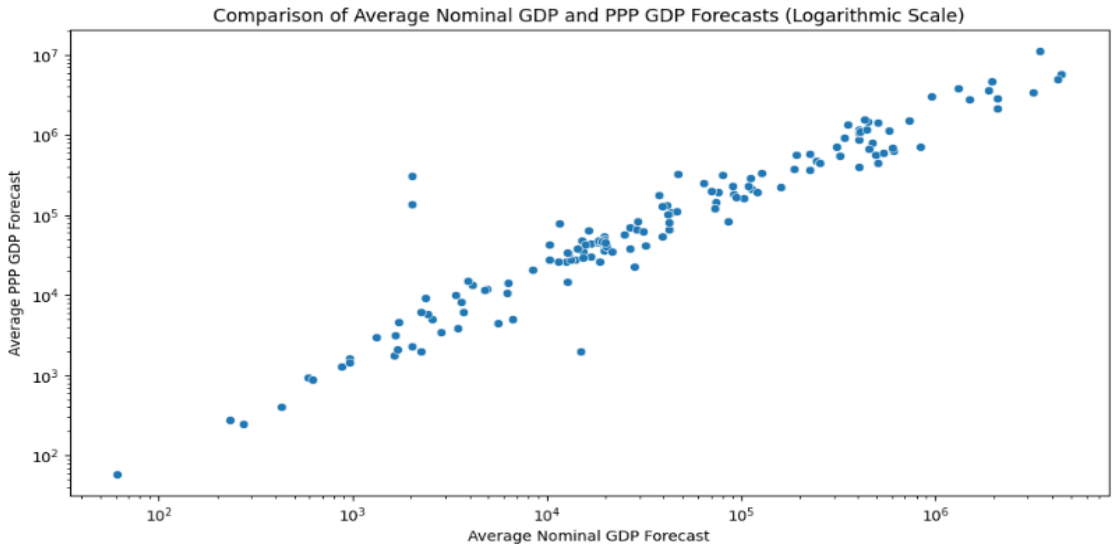
Scatterplot:

```
# Plotting GDP Forecasts
plt.figure(figsize=(12, 6))
sns.scatterplot(x='Average Nominal GDP Forecast', y='Average PPP GDP Forecast', data=df)
plt.title('Comparison of Average Nominal GDP and PPP GDP Forecasts')
plt.xlabel('Average Nominal GDP Forecast')
plt.ylabel('Average PPP GDP Forecast')
```



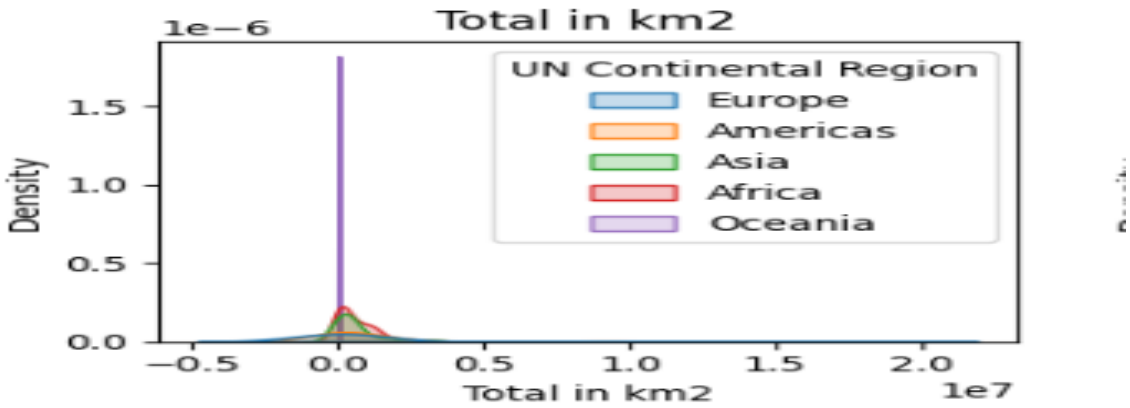
Scatterplot:

```
# Plotting GDP Forecasts with logarithmic scale
plt.figure(figsize=(12, 6))
sns.scatterplot(x='Average Nominal GDP Forecast', y='Average PPP GDP Forecast', data=df)
plt.xscale('log')
plt.yscale('log')
plt.title('Comparison of Average Nominal GDP and PPP GDP Forecasts (Logarithmic Scale)')
plt.xlabel('Average Nominal GDP Forecast')
plt.ylabel('Average PPP GDP Forecast')
plt.show()
```

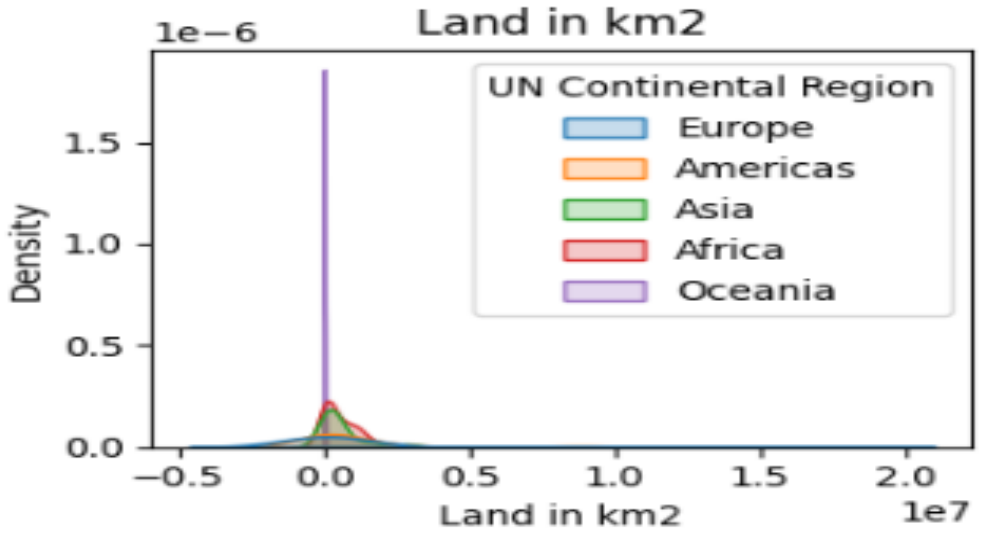


Kde plot:

```
categorical_cols = ['UN Continental Region']
columns = df.columns.tolist()
numerical_cols = [elem for elem in columns if elem not in categorical_cols + ['Country', 'UN Statistical Subregion']]
for category in categorical_cols:
    plt.figure(figsize=(15,15))
    for ax, col in enumerate(numerical_cols):
        plt.subplot(5,4, ax+1)
        plt.title(col)
        sns.kdeplot(x=df[col],shade=True, hue=df[category])

    plt.tight_layout()
```

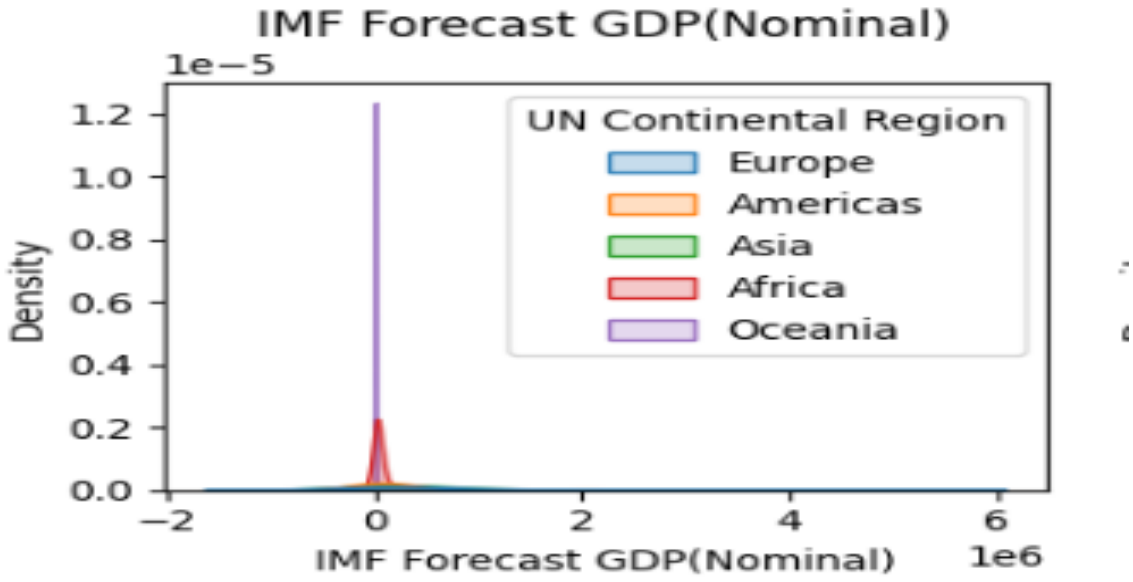


Kde plot- Land in km2:



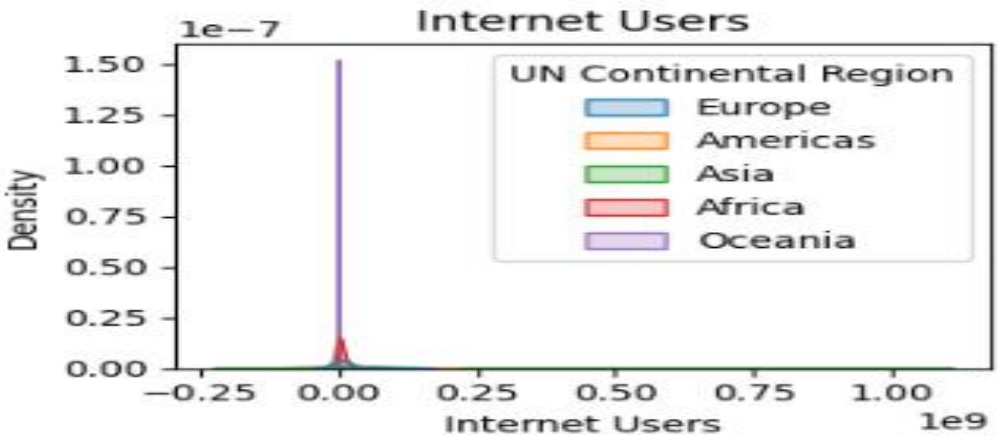
From the above plot it is clear that Oceania occupies more land when compared to other Continents.

Kde plot – IMF Forecast GDP(Nominal):

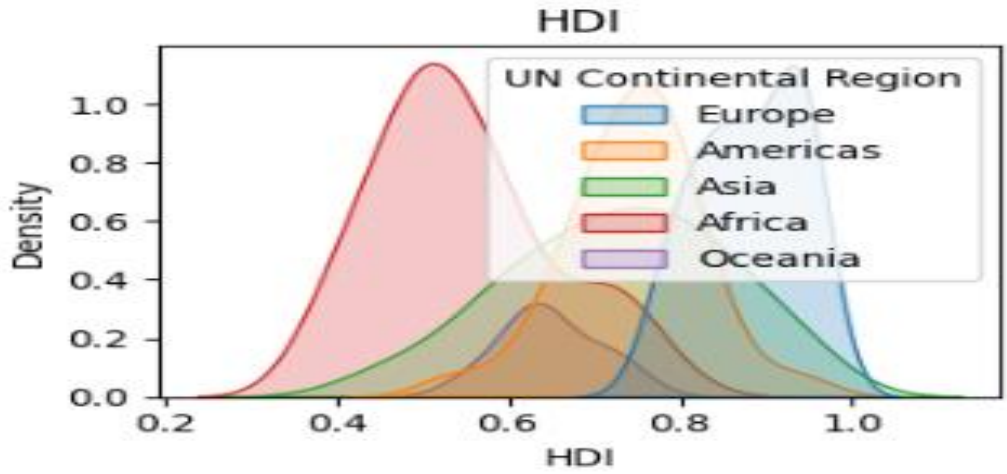


From the above plot, it can be inferred that Oceania has a higher GDP compared to others.

Kde plot of internet users:

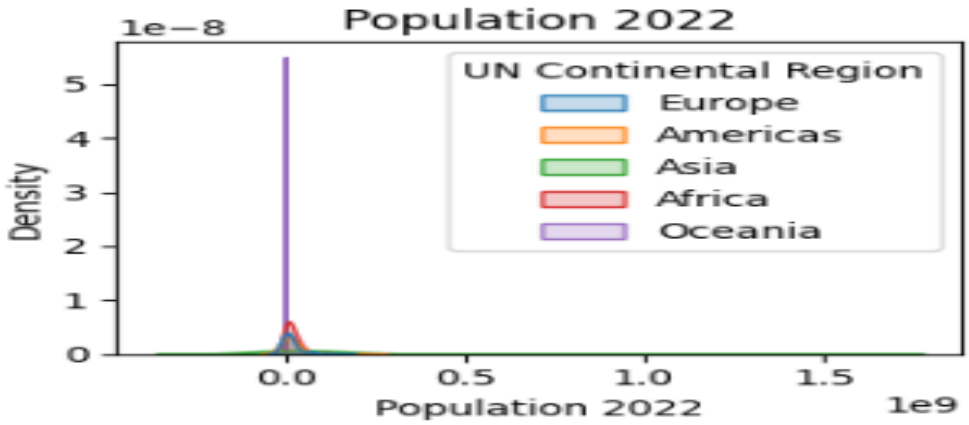


Kde plot of HDI:



It can be inferred that Asia has wide spread of data but slightly skewed to the left.

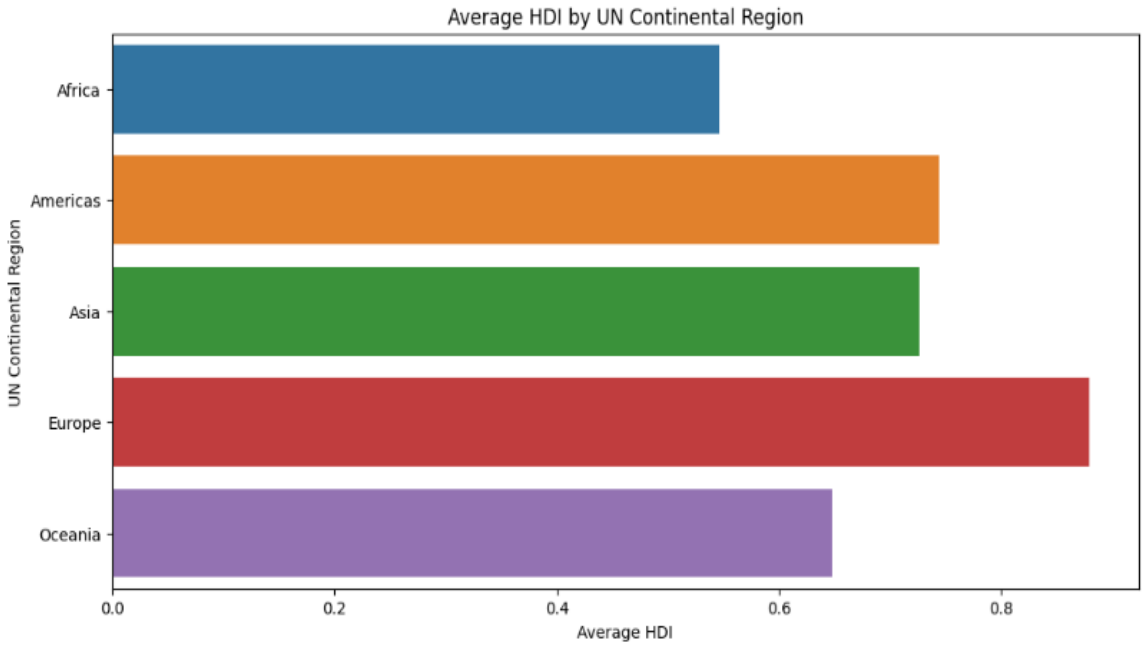
Kde plot of population 2022:



Bar graph:

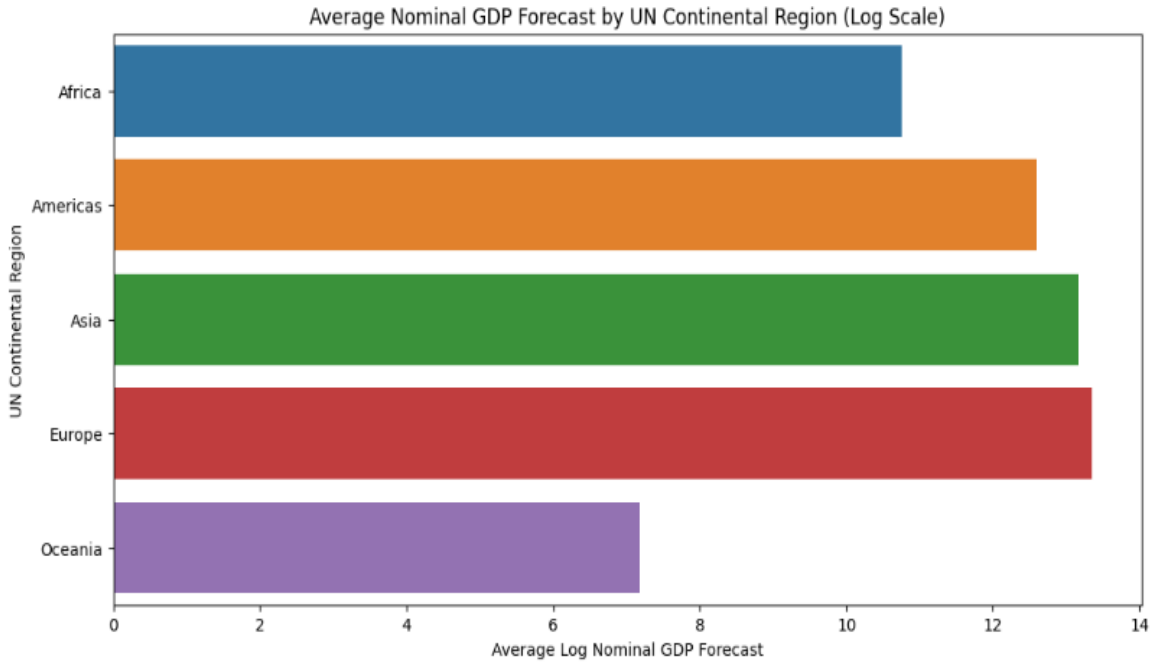
Average HDI by UN continental region:

```
# Set a logarithmic scale for GDP values
region_avg['Log IMF Forecast GDP(Nominal)'] = np.log1p(region_avg['IMF Forecast GDP(Nominal)'])
region_avg['Log IMF Forecast GDP(PPP)'] = np.log1p(region_avg['IMF Forecast GDP(PPP)'])

# Plotting Average HDI by Region
plt.figure(figsize=(12, 6))
sns.barplot(x='HDI', y='UN Continental Region', data=region_avg)
plt.title('Average HDI by UN Continental Region')
plt.xlabel('Average HDI')
plt.ylabel('UN Continental Region')
plt.show()

# Plotting Average Nominal GDP by Region on a logarithmic scale
plt.figure(figsize=(12, 6))
sns.barplot(x='Log IMF Forecast GDP(Nominal)', y='UN Continental Region', data=region_avg)
plt.title('Average Nominal GDP Forecast by UN Continental Region (Log Scale)')
plt.xlabel('Average Log Nominal GDP Forecast')
plt.ylabel('UN Continental Region')
plt.show()
```



It can be concluded from the above plot that Europe has the most HDI out of all the continents.

Log version of nominal GDP forecast by UN continental region:



Above bar graph shows that the Europe has the highest GDP compared to others.

Advantages of seaborn:

- Seaborn requires less code to create complex visualizations compared to Matplotlib.
- It helps you explore and understand your data.
- It provides data visualizations that are typically more aesthetic and statistically sophisticated.
- It provides a high-level API that abstracts away many of the low-level details.
- We may switch to any other data format using the 'kind' parameter within this.

Disadvantages of seaborn:

- It can be slow and memory-intensive for large or complex datasets as it relies on matplotlib as its backend, which is not optimized for performance or scalability.
- It is conceivable that your data will need to be reformatted. It's charting tools are expressive when given a clean long-form dataset.
- There are two sorts of plotting functions. So, we can't mix a seaborn plot with matplotlib figure with several axes.

----------THE END----------