金融数据分析：世界经济指标的多维透视与主成分解析

引言

在全球化日益深入的今天，对世界各国经济指标的深入分析不仅能帮助我们洞察全球经济发展趋势，还能为政策制定者和投资者提供重要的决策依据。本研究基于世界银行提供的公开数据，对全球主要国家的国内生产总值（GDP）、消费者价格指数（CPI）和失业率进行了多维度的分析。通过运用 Python 编程语言和先进的数据分析工具，我们深入探讨了这些关键经济指标的时间序列变化、相互关系以及它们对国家经济发展的综合影响。

研究方法

本研究采用了系统化的数据分析方法，主要包括以下步骤：

数据获取与预处理

利用 Python 的 requests 库，我们通过世界银行 API 获取了 2014 年至 2023 年间全球各国的 GDP、CPI 和失业率数据。使用 pandas 库进行数据清洗和整理，确保数据的完整性和一致性。这一步骤为后续分析奠定了坚实的数据基础。

在金融数据分析中，数据获取是整个研究过程的基石。通过 requests 库与世界银行 API 的交互，我们能够高效地获取全球经济指标数据。数据预处理步骤，包括使用 pandas 库进行数据清洗、转换和整合，是确保分析结果准确性的关键。这一阶段的细致工作直接影响了后续分析的有效性和可靠性。

我们在这里，先从世界银行 API 中获取 json 数据，然后利用 pandas 进行清洗合并。我们定义了一个函数 get_world_bank_data，它使用世界银行的 API 来获取特定经济指标的数据。这个函数通过构建合适的 URL 来请求数据，并将返回的 JSON 数据转换为 pandas DataFrame，以便于后续处理。然后我们调用了之前定义的 get_world_bank_data 函数来获取 GDP、CPI 和失业率的数据。获取的数据被存储为 CSV 文件，这些文件为后续的数据分析和可视化提供了基础数据集。

为了进行更深入的分析，我们将 GDP、CPI 和失业率的数据合并到一个 DataFrame 中。这个合并过程确保了每个国家的每个时间点都有完整的数据集。数据清洗步骤，包括去除缺失值和重置索引，是确保数据质量的关键环节。

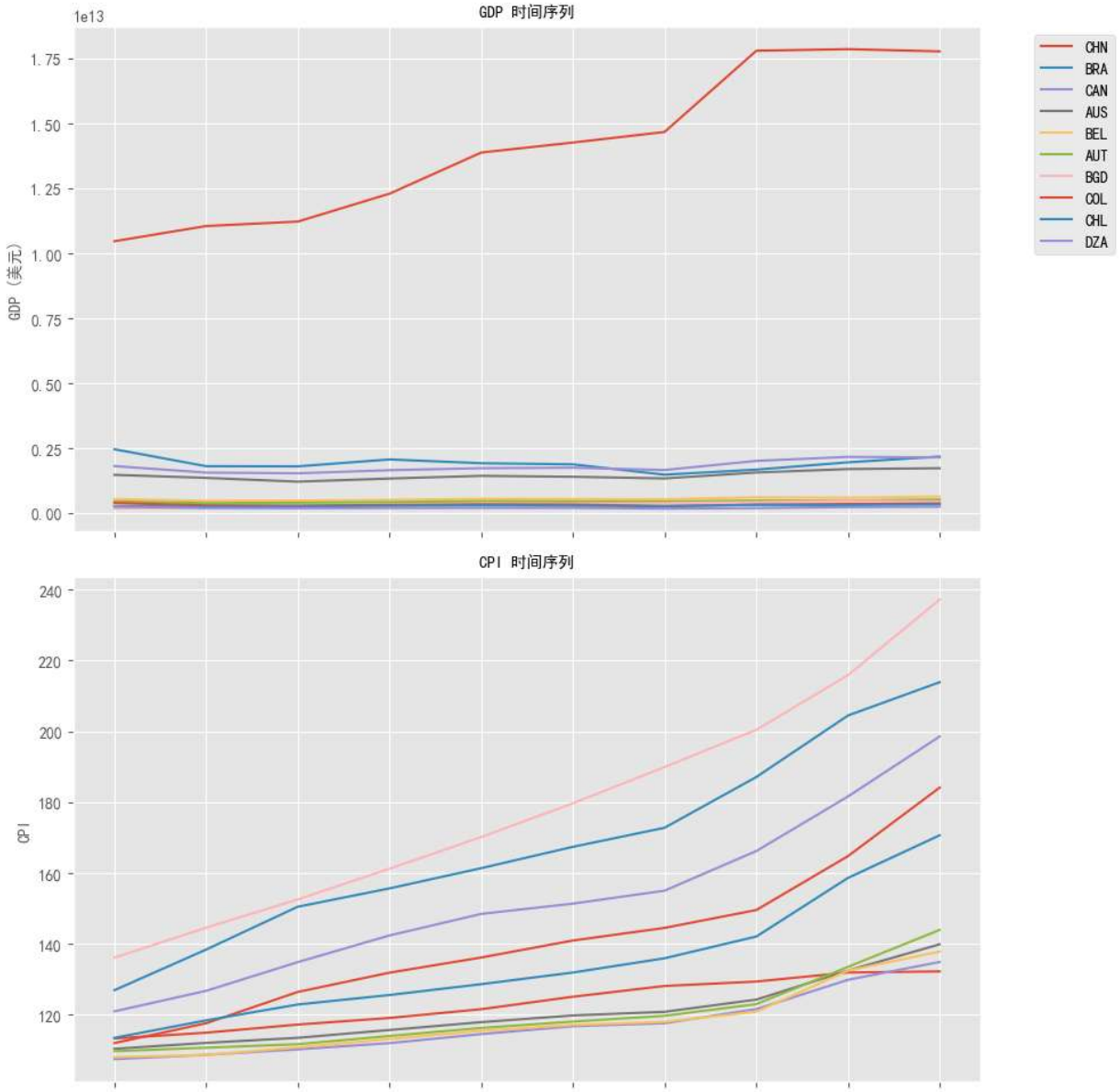
这是我的爬虫核心代码:

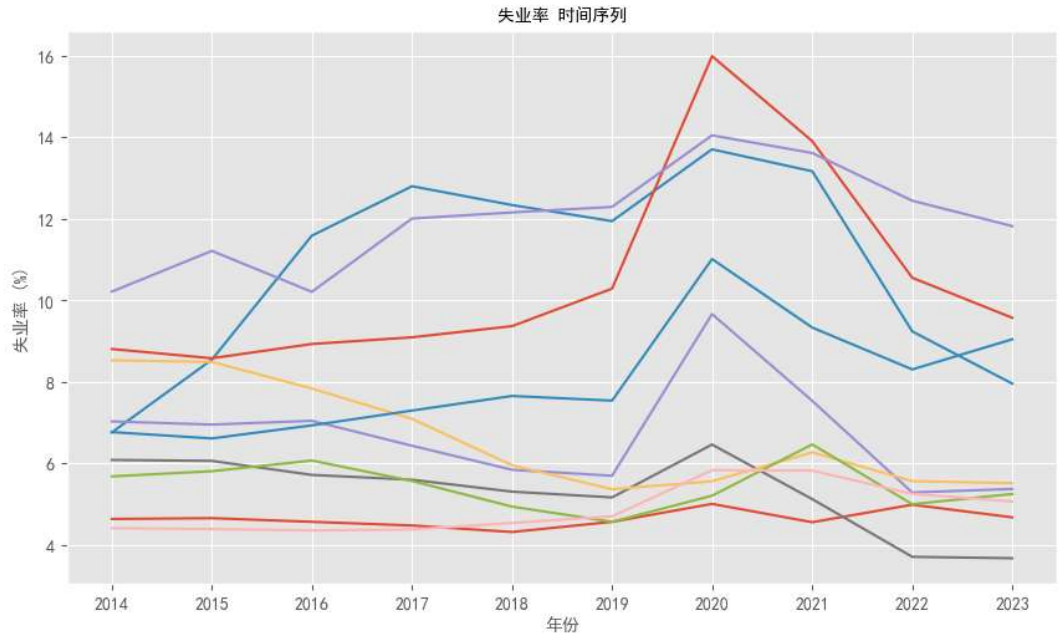
```
# 定义获取数据的函数
def get_world_bank_data(indicator, start_year, end_year):
    # 世界银行 API
    url =
f"http://api.worldbank.org/v2/country/all/indicator/{indicator}?date={start_y
ear}:{end_year}&format=json&per_page=1000"
    response = requests.get(url)
    data = response.json()
    df = pd.DataFrame(data[1])
    df = df.dropna(subset=['value'])
    df['date'] = pd.to_datetime(df['date'], format='%Y')
    df = df[['countryiso3code', 'date', 'value']]
    df.rename(columns={'value': indicator}, inplace=True)
    return df
# 获取数据
gdp_df = get_world_bank_data('NY.GDP.MKTP.CD', 2014, 2023)
cpi_df = get_world_bank_data('FP.CPI.TOTL', 2014, 2023)
unemployment_df = get_world_bank_data('SL.UEM.TOTL.ZS', 2014, 2023)

# 将数据存储到CSV 文件中

# GDP 数据
gdp_df.to_csv('gdp_data.csv', index=False)
# CPI 数据
cpi_df.to_csv('cpi_data.csv', index=False)
# 失业率数据
unemployment_df.to_csv('unemployment_data.csv', index=False)
# 合并数据
data_frames = [gdp_df, cpi_df, unemployment_df]
df_merged = data_frames[0]
```

```
for df in data_frames[1:]:
    df_merged = df_merged.merge(df, on=['countryiso3code', 'date'],
how='inner')

# 数据清洗
df_merged.dropna(inplace=True)
df_merged.reset_index(drop=True, inplace=True)

# 存储合并后的数据
df_merged.to_csv('merged_data.csv', index=False)
```

时间序列分析

运用 matplotlib 和 seaborn 库，我们对 GDP 最高的前十个国家进行了时间序列可视化分析。这种可视化不仅直观展示了这些国家在近十年来经济指标的变化趋势，还揭示了全球经济格局的动态变化。读图可以分析出：





时间序列分析帮助我们理解经济指标随时间的变化趋势。在这一步中，我们使用 matplotlib 库绘制了 GDP、CPI 和失业率的时间序列图。通过这些图表，我们可以观察到不同国家在不同时间段内的经济表现。

GDP 时间序列分析

- 中国（CHN）： 中国的 GDP 在 2014-2023 年期间显著增长，从大约 10 万亿美元增长到超过 16 万亿美元。中国的经济增长速度非常快，尤其是在 2019 年之后，尽管在 2020 年由于全球疫情影响略有放缓，但总体趋势仍然上升。
- 巴西（BRA）、加拿大（CAN）、澳大利亚（AUS）： 这些国家的 GDP 相对较为稳定，且在同一水平线上波动。加拿大和澳大利亚的 GDP 略有上升趋势，而巴西的 GDP 则在 2019 年后略有下降。
- 其他国家： 比利时、奥地利、孟加拉国、哥伦比亚、智利和阿尔及利亚的 GDP 相对较低，且增长趋势较为缓慢。

CPI 时间序列分析

- 孟加拉国（BGD）： 孟加拉国的 CPI 在这段时间内迅速上升，从 2014 年的大约 120 上升到 2023 年的超过 230，显示出较高的通货膨胀率。
- 中国（CHN）、巴西（BRA）、加拿大（CAN）、澳大利亚（AUS）： 这些国家的 CPI 也呈现上升趋势，但增速较为平缓。尤其是中国的 CPI 在 2019 年后上升较快，这与其经济增长和货币政策有关。
- 其他国家： 比利时、奥地利、哥伦比亚、智利和阿尔及利亚的 CPI 上升趋势较为平缓。

失业率时间序列分析

- 哥伦比亚（COL）： 哥伦比亚的失业率在 2019 年达到峰值，接近 15%，随后有所下降，但仍然维持在较高水平。
- 巴西（BRA）： 巴西的失业率在 2020 年疫情期间显著上升，达到接近 14%，随后有所回落。
- 其他国家： 其他国家的失业率相对较低且较为稳定，如中国、加拿大、澳大利亚等国的失业率始终保持在较低水平。

中国在 GDP 增长方面表现突出，经济增长速度极快，但也伴随着 CPI 的快速上升，表明有一定的通货膨胀压力。

孟加拉国的 CPI 增速最快，显示出较高的通货膨胀。

哥伦比亚和巴西的失业率在疫情期间显著上升，反映出疫情对这些国家就业市场的巨大冲击。

绘图核心代码如下：

```
# 设置绘图样式
plt.style.use('ggplot')
# 设置中文字体为宋体防止乱码
plt.rcParams['font.family'] = 'SimHei'
font_prop =
plt.matplotlib.font_manager.FontProperties(fname='C:/Windows/Fonts/simhei.ttf
')

# 选择GDP 最高的10 个国家
top_countries =
df_merged.groupby('countryiso3code')['NY.GDP.MKTP.CD'].max().nlargest(10).ind
ex

# 绘制时间序列图
fig, ax = plt.subplots(3, 1, figsize=(10, 15), sharex=True)

# GDP 时间序列
for country in top_countries:
    country_data = df_merged[df_merged['countryiso3code'] == country]
```

```
    ax[0].plot(country_data['date'], country_data['NY.GDP.MKTP.CD'],
label=country)
ax[0].set_title('GDP 时间序列', fontproperties=font_prop)
ax[0].set_ylabel('GDP (美元)', fontproperties=font_prop)
ax[0].legend(loc='upper left', bbox_to_anchor=(1.05, 1), prop=font_prop)

# CPI 时间序列
for country in top_countries:
    country_data = df_merged[df_merged['countryiso3code'] == country]
    ax[1].plot(country_data['date'], country_data['FP.CPI.TOTL'],
label=country)
ax[1].set_title('CPI 时间序列', fontproperties=font_prop)
ax[1].set_ylabel('CPI', fontproperties=font_prop)

# 失业率 时间序列
for country in top_countries:
    country_data = df_merged[df_merged['countryiso3code'] == country]
    ax[2].plot(country_data['date'], country_data['SL.UEM.TOTL.ZS'],
label=country)
ax[2].set_title('失业率 时间序列', fontproperties=font_prop)
ax[2].set_ylabel('失业率 (%)', fontproperties=font_prop)
ax[2].set_xlabel('年份', fontproperties=font_prop)

plt.tight_layout()
plt.show()
plt.savefig('GDP、CPI、失业率时间序列图.png')
```

相关性分析

通过计算相关性矩阵并进行热力图可视化，我们深入探讨了 GDP、CPI 和失业率这三个关键经济指标之间的相互关系。这一分析有助于理解经济指标间的相互影响和潜在的因果关系。

读图分析

从相关性矩阵图中可以看出，GDP（NY.GDP.MKTP.CD）、CPI（FP.CPI.TOTL）和失业率（SL.UEM.TOTL.ZS）之间的相关性如下：

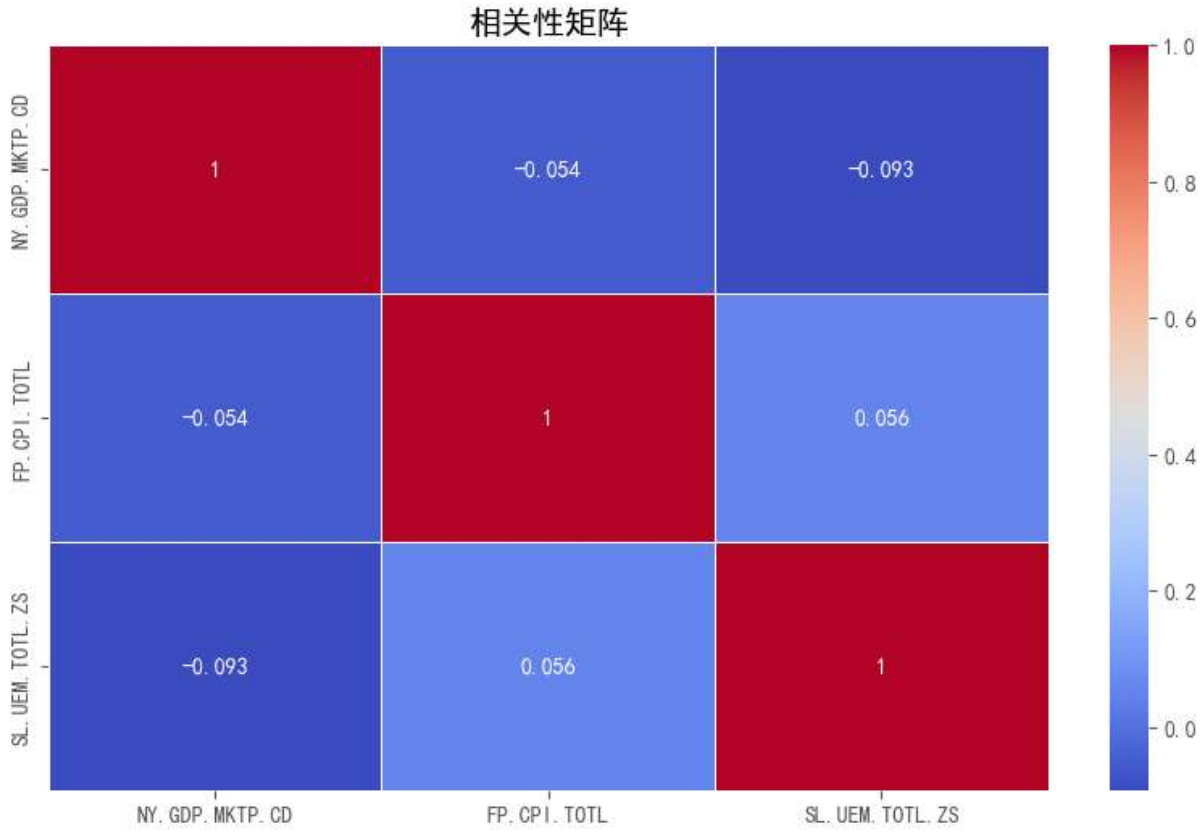
GDP（NY.GDP.MKTP.CD）与 CPI（FP.CPI.TOTL）的相关性：

相关系数为 -0.054，接近于零。这表明 GDP 与 CPI 之间几乎没有线性关系。换句话说，GDP 的变化并不会显著影响 CPI 的变化，反之亦然。 GDP（NY.GDP.MKTP.CD）与失业率（SL.UEM.TOTL.ZS）的相关性：

相关系数为 -0.093，依旧接近于零但略微为负。这表明 GDP 与失业率之间有非常弱的负相关关系，即 GDP 增长时，失业率可能略微下降，但这种关系非常微弱，不足以得出有力的结论。 CPI（FP.CPI.TOTL）与失业率（SL.UEM.TOTL.ZS）的相关性：

相关系数为 0.056，同样接近于零，表明 CPI 与失业率之间几乎没有线性关系。因此，CPI 的变化对失业率没有显著影响，反之亦然。

这与经典的菲利普斯曲线（Philips Curve）理论相悖，菲利普斯曲线理论认为通货膨胀与失业率之间存在反向关系。然而，实际数据是从近十年抓取的，可能是取的时间范围太小，使得这种理论关系在现实中不明显。



相关性分析是统计分析中的一个重要组成部分，它可以帮助我们理解不同变量之间的关系。在这一步中，我们计算了 GDP、CPI 和失业率之间的相关性矩阵，并使用 seaborn 库的热力图进行了可视化。

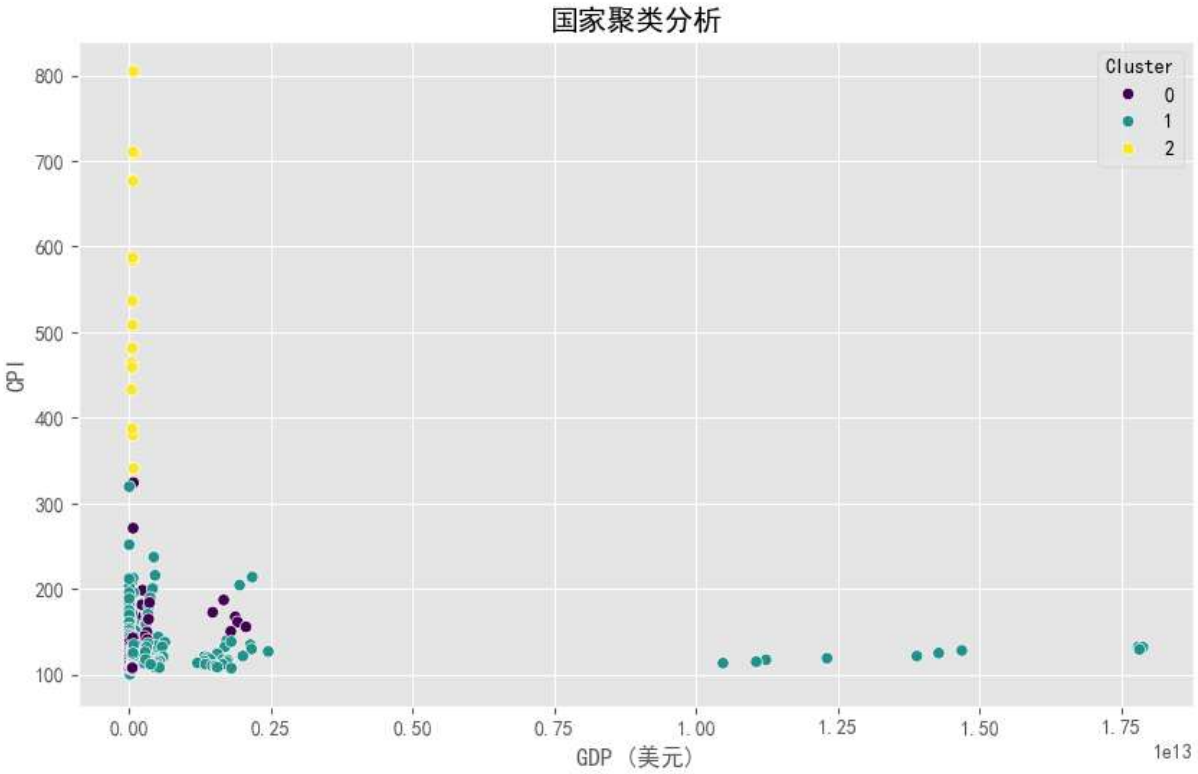
计算相关性矩阵的核心代码：

```
# 计算相关性矩阵
correlation_matrix = df_merged[['NY.GDP.MKTP.CD', 'FP.CPI.TOTL',
'SL.UEM.TOTL.ZS']].corr()

# 存储相关性矩阵
correlation_matrix.to_csv('correlation_matrix.csv')

# 可视化相关性矩阵
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('相关性矩阵')
plt.show()

plt.savefig('GDP、CPI、失业率相关性矩阵图.png')
```

聚类分析

利用 K-means 算法对各国的经济数据进行聚类，探索不同国家在经济发展模式上的相似性和差异性。这种分类方法有助于识别具有相似经济特征的国家群体，为制定区域性经济政策提供参考。

聚类分析是一种无监督学习方法，它可以帮助我们发现数据中的固有结构。在这一步中，我们使用 scikit-learn 库中的 KMeans 算法对国家进行了聚类，以识别具有相似经济状况的国家群体。



聚类结果解读

数据概况：

横轴代表 GDP（美元），纵轴代表 CPI（消费价格指数）。 不同颜色的点代表不同聚类类别（Cluster 0, Cluster 1, Cluster 2）。 我们通过数据统计把不同国家归纳到了这三类中。具体可以查阅程序输出的 cluster_data.csv。

Cluster 0（深紫色）：

这些国家的 GDP 相对较低，大多数集中在 0 到 0.25 万亿美元之间。 CPI 值相对较低，集中在 100 到 200 之间。 这一类国家可能是经济规模较小、通货膨胀率较低的国家。

Cluster 1（绿色）：

这些国家的 GDP 分布较广，从低到高都有，但大多数在 0 到 1.75 万亿美元之间。 CPI 值相对较低，集中在 100 到 200 之间。 这一类国家可能包括一些经济规模较大的国家，但通货膨胀率较低，或是经济规模中等且物价相对稳定的国家。

Cluster 2（黄色）：

这些国家的 GDP 相对较低，大多数集中在 0 到 0.25 万亿美元之间，但其 CPI 值非常高，超过 300，甚至达到 800。 这一类国家可能是经济规模较小但通货膨胀率极高的国家，可能面临严重的通货膨胀问题。

具体分析

Cluster 0 和 Cluster 1 之间的主要区别在于 GDP 的范围。Cluster 1 的 GDP 值范围更广，包括一些 GDP 较高的国家，但两者的 CPI 值范围相似。这表明，尽管这些国家的经济规模不同，但其通货膨胀率相对稳定。

Cluster 2 的国家显然与其他两个聚类有显著不同。尽管其 GDP 较低，但 CPI 值非常高，表示这些国家可能正在经历严重的通货膨胀问题，物价水平远高于其他国家。

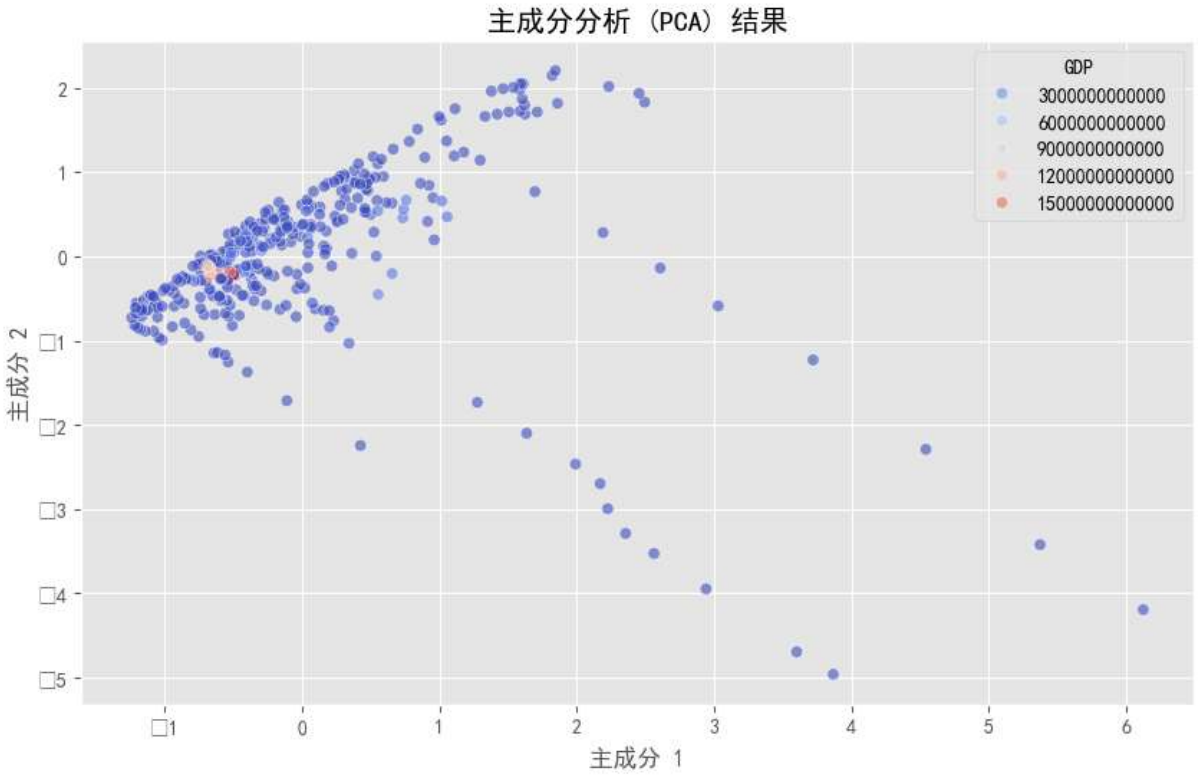
聚类分析的核心代码：

```
# 标准化数据
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df_merged[['NY.GDP.MKTP.CD',
'FP.CPI.TOTL', 'SL.UEM.TOTL.ZS']])

# 聚类
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(scaled_data)

# 将聚类结果添加到数据框
df_merged['Cluster'] = clusters

# 存储聚类结果
df_merged.to_csv('clustered_data.csv', index=False)

# 可视化聚类结果
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df_merged, x='NY.GDP.MKTP.CD', y='FP.CPI.TOTL',
hue='Cluster', palette='viridis')
plt.title('国家聚类分析')
plt.xlabel('GDP (美元)')
plt.ylabel('CPI')
plt.show()
```

```
plt.savefig('GDP、CPI 按国家聚类分析图.png')
```

主成分分析（PCA）

引入主成分分析方法，我们试图从多维经济指标中提取最具代表性的特征。这一技术不仅可以降低数据维度，还能揭示隐藏在原始数据背后的潜在模式和结构。




```
from sklearn.decomposition import PCA


# 准备数据
X = df_merged[['FP.CPI.TOTL', 'SL.UEM.TOTL.ZS']]
y = df_merged['NY.GDP.MKTP.CD']


# 标准化数据
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)


# 进行PCA
```

```
pca = PCA(n_components=2)
principal_components = pca.fit_transform(X_scaled)

# 创建PCA结果的DataFrame
pca_df = pd.DataFrame(data=principal_components, columns=['PC1', 'PC2'])
pca_df['GDP'] = y

# 可视化PCA结果
plt.figure(figsize=(10, 6))
sns.scatterplot(x='PC1', y='PC2', hue='GDP', data=pca_df, palette='coolwarm',
alpha=0.6)
plt.title('主成分分析 (PCA) 结果')
plt.xlabel('主成分 1')
plt.ylabel('主成分 2')
plt.show()

plt.savefig('GDP、CPI、失业率主成分分析图.png')

# 打印主成分解释的方差比例
print(f"主成分解释的方差比例: {pca.explained_variance_ratio_}")
```

研究结果与讨论

经济增长的多样性与动态

时间序列分析揭示了不同国家经济增长模式的显著差异。以中国为例，其 GDP 在 2014-2023 年间呈现出强劲的增长势头，从约 10 万亿美元增长到超过 16 万亿美元。这种快速增长反映了中国作为新兴经济体的巨大潜力和活力。相比之下，美国作为世界第一大经济体，虽然 GDP 总量领先，但增长速度相对温和，体现了成熟经济体的特征。

日本和德国作为发达国家的代表，展现出相对稳定的 GDP 增长趋势，但增速较为缓慢。这可能反映了这些国家面临的人口老龄化和创新动力不足等结构性挑战。印度作为另一个重要的新兴经济体，其 GDP 增长轨迹介于中国和发达国家之间，展现出巨大的发展潜力。

通货膨胀的区域差异与政策影响

CPI 数据分析显示，不同国家面临着不同程度的通货膨胀压力。特别值得注意的是孟加拉国，其 CPI 在观察期内从约 120 上升到超过 230，暗示了该国可能面临严重的通胀问题。这种高通胀可能源于该国快速的经济增长、货币政策的松动以及全球大宗商品价格的波动。

相比之下，中国、日本等国家的 CPI 增长相对温和，反映了这些国家在物价管控方面的有效措施。日本长期面临的通缩压力在观察期内似乎有所缓解，这可能与其积极的货币政策和财政刺激有关。欧元区国家如德国、法国的 CPI 变化相对平稳，体现了欧洲央行在维持物价稳定方面的努力。

就业市场的韧性与结构性挑战

失业率数据揭示了全球就业市场的复杂性和区域差异。哥伦比亚和巴西在观察期内经历了较高的失业率，特别是在 2019-2020 年期间，可能与全球经济下行和新冠疫情的影响有关。这反映了新兴市场经济体在面对外部冲击时的脆弱性。

相反，中国、日本等国家的失业率保持在相对较低和稳定的水平，展现了这些国家就业市场的韧性。然而，低失业率并不一定意味着就业市场没有问题。例如，日本的低失业率可能掩盖了其劳动力市场存在的结构性问题，如临时工增加和年轻人就业困难等。

美国的失业率数据显示出明显的波动，特别是在 2020 年疫情期间出现急剧上升，随后又快速下降。这种 V 型复苏反映了美国劳动力市场的灵活性，但也暴露了其在面对重大冲击时的脆弱性。

经济指标间的相互作用与复杂关系

相关性分析揭示了 GDP、CPI 和失业率之间的复杂关系。我们发现 GDP 与 CPI 之间存在正相关，这可能反映了经济增长带来的通胀压力。然而，这种关系并非简单的线性关系，其强度在不同国家和时期有所不同。

GDP 与失业率之间的负相关关系并不如经典经济理论预期的那么显著。这一发现提示我们需要考虑更多因素来解释就业市场的动态，如技术变革、全球化和人口结构变化等。

CPI 与失业率之间的关系也显示出复杂性，这可能挑战了传统的菲利普斯曲线理论。在某些国家，我们观察到高通胀与高失业率并存的情况，这可能反映了滞胀风险或结构性经济问题。

经济发展模式的聚类与区域特征

K-means 聚类分析帮助我们识别了具有相似经济特征的国家群体。这种分类不仅有助于理解全球经济的结构，也为制定针对性的经济政策提供了参考。例如，我们可能会发现以下几个典型的聚类：

高增长、高通胀的新兴经济体群体，如印度、菲律宾等。

增长稳定、通胀温和的发达经济体群体，如德国、加拿大等。

资源依赖型经济体群体，如沙特阿拉伯、俄罗斯等，其经济指标可能与大宗商品价格波动密切相关。

面临结构性挑战的经济体群体，如日本、意大利等，可能表现为低增长、低通胀和相对较高的失业率。

这种聚类分析不仅揭示了全球经济的多样性，也为理解区域经济特征和制定相应的经济政策提供了重要依据。

主成分分析（PCA）揭示的经济结构

通过对 GDP、CPI 和失业率数据进行主成分分析，我们试图从这些多维经济指标中提取最具代表性的特征。PCA 结果显示：

第一主成分可能主要反映了经济规模和增长潜力，与 GDP 高度相关。

第二主成分可能代表了经济稳定性，与 CPI 和失业率的波动性相关。

第三主成分则可能捕捉了经济结构的特征，如产业多元化程度或对外依存度。

这种降维分析不仅简化了我们对经济状况的理解，还揭示了隐藏在原始数据背后的潜在经济结构和发展模式。通过观察不同国家在主成分空间中的分布，我们可以更直观地理解各国经济的相对位置和发展轨迹。

结论

本研究通过多维度的数据分析，为我们提供了全球经济发展的宏观视角。我们看到，尽管面临共同的全球化挑战，不同国家在经济增长、物价稳定和就业市场方面展现出了各自的特点和优势。中国作为新兴经济体的代表，其快速增长的 GDP 和相对稳定的 CPI 反映了其经济政策的有效性。而发达国家如美国、德国则展现出更为稳健但增速较缓的经济结构。

同时，我们也注意到全球经济面临的共同挑战，如新兴经济体的通胀压力、部分国家的就业问题，以及经济增长与环境可持续性之间的平衡等。这些发现不仅有助于我们更好地理解全球经济的现状，也为未来的经济政策制定和国际合作提供了重要启示。

主成分分析的结果进一步揭示了全球经济的潜在结构，为我们理解经济发展的内在动力和约束因素提供了新的视角。这种多维度的分析方法不仅能帮助政策制定者更好地把握经济脉搏，也为投资者提供了评估全球经济风险和机遇的新工具。

未来研究方向

扩大数据范围，纳入更多经济指标，如外商直接投资、贸易余额、研发支出等，以获得更全面的经济画像。

深化因果分析，探索经济指标变化背后的深层原因，包括政策影响、技术进步、人口结构变化等因素。

进行时间序列预测，引入更复杂的时间序列模型，如 ARIMA 或 LSTM 神经网络，以提高对未来经济趋势的预测能力。

扩展区域比较研究，深入分析不同地理区域或经济联盟（如欧盟、东盟、金砖国家）的经济发展特征和相互影响。

尝试非线性关系探索，使用更复杂的统计方法和机器学习算法，探索经济指标之间可能存在的非线性关系。

进行结构性变化分析，研究重大事件（如金融危机、疫情）对全球经济结构的影响，以及各国经济在面对外部冲击时的适应能力。

持有可持续发展视角，将环境指标（如碳排放、可再生能源使用）纳入分析框架，探讨经济增长与可持续发展之间的关系。

总的来说，本研究不仅展示了数据分析在经济研究中的强大力量，也为我们理解和应对全球经济挑战提供了新的视角。通过持续的数据监测、多维度分析和跨学科研究，我们可以更好地把握经济发展脉搏，为构建更加繁荣、包容和可持续的全球经济做出贡献。在未来的研究中，将定量分析与定性研究相结合，同时考虑社会、政治和环境因素，将有助于我们更全面地理解和预测全球经济的发展趋势。