Análise dos Índices de Corrupção e Desenvolvimento Humano

```
# Importar bibliotecas
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats

# Carregar arqivo CPI2023.csv
df = pd.read_csv('CPI2023.csv')

# Mostra o cabeçalho do dataset
print(df.head(10))

# Mostrar as informações do dataset
print(df.info())
```

Ajuste de Dados:

Limpeza dos dados: Identifique e trate valores ausentes, valores duplicados e valores discrepantes (outliers). Documente as etapas e decisões tomadas.

Transformação de dados: Realize transformações necessárias, como normalização ou padronização de variáveis numéricas, codificação de variáveis categóricas e criação de novas variáveis derivadas, se aplicável.

```


# Renoemar as colunas
df.columns = ['Country', 'ISO3', 'Region', 'CPI', 'Rank', 'SE',
'Sources', 'Lower_CI', 'Upper_CI', 'ADB_CPIA', 'Bertelsmann_SGI',
'Bertelsmann_TII', 'EIU_Country_Ratings', 'Freedom_House',
'Global_Insights', 'IMD_WCY', 'PERC_ARG', 'PRS_ICRG', 'V_Democracy',
'WB_CPIA', 'WEF_EOS', 'WJP_Rule_Law']

## Limpeza do dataset para remover dados desnecessários para a análise
# Remover colunas numericas onde tem mais de 100 valores nulos
df = df.dropna(thresh=100, axis=1)

# Pegar apenas as colunas numericas, necessário para algumas análises
df_num = df.select_dtypes(include=[np.number])

# Verificar a correlação entre as variáveis para limpeza do dataset,
isso será feito para remover variáveis que possuem pouco ou nenhuma
```

```
relação com o índice principal (CPI)
correlation = df_num.corr(method='pearson')
sns.heatmap(correlation, annot=True, cmap='coolwarm')
plt.title('Correlação entre as variáveis do dataset')
plt.show()
plt.clf()

# Remover colunas que possuam pouca relação com a variável principal
(CPI)
df_num = df_num.drop(['SE', 'Sources', 'Lower_CI', 'Upper_CI',
'WJP_Rule_Law'], axis=1)
print(df.info())
```

Descrição de cada coluna

- Country: Nome do país
- ISO3: Código ISO3 do país
- Region: Região do país
- CPI: Índice de Percepção de Corrupção (Corruption Perception Index)
- Rank: Ranking do CPI
- Bertelsmann_TII: Bertelsmann Stiftung - Transformation Index (Índice de Transformação)
- EIU_Country_Ratings: Economist Intelligence Unit - Country Ratings (Avaliações de Países)
- Global_Insights: Global Insights (Percepção Global)
- PRS_ICRG: PRS International Country Risk Guide (Índice de Risco País)
- V_Democracy: V-Dem Institute (Varieties of Democracy)

Explicações dos Termos:

- Bertelsmann_TII: Bertelsmann Stiftung - Transformation Index (Índice de Transformação)
  - Descrição: Também criado pela Fundação Bertelsmann, este índice avalia a transformação de países em desenvolvimento e em transição em direção a uma democracia e uma economia de mercado sustentáveis.
  - Objetivo: Medir o progresso dos países em termos de transformação política e econômica, bem como a gestão política.
- EIU_Country_Ratings: Economist Intelligence Unit - Country Ratings (Avaliações de Países)
  - Descrição: As avaliações da Economist Intelligence Unit fornecem uma análise detalhada das condições políticas, econômicas e sociais de diversos países.
  - Objetivo: Oferecer insights e previsões para empresas, governos e organizações sobre o ambiente de negócios e riscos em diferentes países.
- Global_Insights: Global Insights (Percepção Global)

  - Descrição: Este índice mede a percepção global sobre diferentes aspectos políticos e econômicos de diversos países.
  - Objetivo: Oferecer uma perspectiva abrangente sobre os riscos e oportunidades globais, ajudando na tomada de decisões informadas.
- PRS_ICRG: PRS International Country Risk Guide (Índice de Risco País)
  - Descrição: O guia de risco país da PRS (Political Risk Services) fornece avaliações de risco político, econômico e financeiro de países ao redor do mundo.
  - Objetivo: Ajudar empresas, investidores e governos a entender e gerenciar riscos país.
- V_Democracy: V-Dem Institute (Varieties of Democracy)
  - Descrição: O Instituto V-Dem oferece um conjunto de índices detalhados que medem diferentes aspectos da democracia em países ao redor do mundo.
  - Objetivo: Fornecer dados e análises sobre a variação e evolução da democracia, ajudando na pesquisa e formulação de políticas.

Análise Exploratória:

Descrição do Dataset:

- Descreva o dataset escolhido: O Índice de Percepção da Corrupção (CPI) de 2023 indica que a corrupção continua sendo um problema generalizado em todo o mundo. Este índice avalia 180 países e territórios com base em como seus setores públicos são percebidos em termos de corrupção, utilizando uma escala de 0 (altamente corrupto) a 100 (muito limpo).

- Fonte: https://www.kaggle.com/datasets/agungpambudi/global-corruption-index-transparency-perceptions

- Contexto: O dataset fornece dados sobre o Índice de Percepção de Corrupção (CPI) de 2023 para diversos países e territórios ao redor do mundo. O CPI é uma medida amplamente utilizada para avaliar o nível percebido de corrupção no setor público de um país, onde pontuações mais altas indicam menores níveis de corrupção percebida. A corrupção é um problema crítico que afeta a governança, a economia e a qualidade de vida das pessoas. Este índice é utilizado por governos, ONGs, empresas e cidadãos para entender e combater a corrupção.

- Objetivo: O objetivo principal do dataset é demonstrar que uma menor percepção de corrupção (maiores pontuações no CPI) está associada a uma melhor qualidade de vida e governança em um país. O dataset permite a análise de como a corrupção afeta diversos aspectos de um país, incluindo governança, risco político, competitividade econômica e o estado de direito.

- Observações (linhas): São 180 observações, cada uma representando um país ou território

- Variáveis (colunas): São 22 variáveis (ao todo), porém ajustei o dataframe para 10 colunas onde removi as que mais tinhas valures nulos e as colunas com menos correlação entre as principais CPI e Rank. Dentre elas temos o nome do país, a pontuação do CPI (Corruption Perception Index), o ranking do CPI, e diversos índices e avaliações de diferentes organizações, como o Banco Mundial, o Fórum Econômico Mundial, e outros, que contribuem para a pontuação do CPI e fornecem uma visão mais ampla sobre a governança e o risco de corrupção nos países
- Resumo estatístico: Calcule e apresente estatísticas descritivas das variáveis numéricas (média, mediana, desvio padrão, mínimo, máximo, quartis) e distribuições de frequência das variáveis categóricas.

```
# Estatísticas descritivas das variáveis numéricas
print('Estatísticas descritivas das variáveis numéricas:')
print(df_num.describe())

# Média das variáveis numéricas
print('Média das variáveis numéricas:')
print(df_num.mean())

# Mediana das variáveis numéricas
print('Mediana das variáveis numéricas:')
print(df_num.median())

# Desvio padrão das variáveis numéricas
print('Desvio padrão das variáveis numéricas:')
print(df_num.std())

# Mínimo das variáveis numéricas
print('Mínimo das variáveis numéricas:')
print(df_num.min())

# Máximo das variáveis numéricas
print('Máximo das variáveis numéricas:')
print(df_num.max())

# Quartis das variáveis numéricas
print('Quartis das variáveis numéricas:')
print(df_num.quantile([0.25, 0.5, 0.75]))

# Covariância entre as variáveis numéricas
cov = df_num.cov()
print('Covariância entre as variáveis numéricas:')
print(cov)

# Distribuições de frequência das variáveis categóricas
print('Distribuição de frequência da variável Region:')
print(df['Region'].value_counts())

# Correlação entre as veriáveis Global_Insights e V_Democracy
```

```
corr_gi_wb = stats.pearsonr(df['Global_Insights'], df['V_Democracy'])
print('Correlação entre as veriáveis Global_Insights e V_Democracy:')
print(corr_gi_wb)
```

Visualizações

Crie visualizações que ajudem a entender a distribuição dos dados e as relações entre as variáveis. Utilize gráficos como histograma, boxplot, scatter plot, heatmap de correlação, entre outros.

```
# Verificar a correlação entre as variáveis ajustadas
correlation = df_num.corr(method='pearson')
sns.heatmap(correlation, annot=True, cmap='coolwarm')
plt.title('Correlação entre as variáveis do dataset')
plt.show()
plt.clf()

# Distribuição de frequência da variável CPI
sns.histplot(df['CPI'], bins=20, kde=True, color='blue')
plt.title('Distribuição de Frequência do CPI (Corruption Perception
Index)')
plt.xlabel('CPI')
plt.ylabel('Frequência')
plt.show()
plt.clf()

# Distribuição de frequência da variável Global_Insights
sns.histplot(df['Global_Insights'], bins=20, kde=True, color='green')
plt.title('Distribuição de Frequência do Global_Insights')
plt.xlabel('Global_Insights')
plt.ylabel('Frequência')
plt.show()
plt.clf()

# Distribuição de frequência da variável V_Democracy
sns.histplot(df['V_Democracy'], bins=20, kde=True, color='red')
plt.title('Distribuição de Frequência do V_Democracy')
plt.xlabel('V_Democracy')
plt.ylabel('Frequência')
plt.show()
plt.clf()

# Lista o Top 10 dos países com maior CPI
sns.barplot(x = df['CPI'], y = df['Country'], order=df.nlargest(10,
'CPI')['Country'], data=df, palette='viridis')
plt.title('Top 10 Países com Maior CPI')
plt.xlabel('CPI')
plt.ylabel('País')
```

```
plt.show()
plt.clf()

# Lista o Top 10 dos países com menor CPI
sns.barplot(x = df['CPI'], y = df['Country'], order=df.nsmallest(10,
'CPI')['Country'], data=df, palette='magma')
plt.title('Top 10 Países com Menor CPI')
plt.xlabel('CPI')
plt.ylabel('País')
plt.show()
plt.clf()

# Lista com Top 10 dos países com maior V_Democracy
sns.barplot(x = df['V_Democracy'], y = df['Country'],
order=df.nlargest(10, 'V_Democracy')['Country'], data=df,
palette='viridis')
plt.title('Top 10 Países com Maior V_Democracy')
plt.xlabel('V_Democracy')
plt.ylabel('País')
plt.show()
plt.clf()

# Lista com Top 10 dos países com menor V_Democracy
sns.barplot(x = df['V_Democracy'], y = df['Country'],
order=df.nsmallest(10, 'V_Democracy')['Country'], data=df,
palette='magma')
plt.title('Top 10 Países com Menor V_Democracy')
plt.xlabel('V_Democracy')
plt.ylabel('País')
plt.show()
plt.clf()

# Scatter plot entre as variáveis CPI e Rank
sns.scatterplot(x = df['CPI'], y = df['Rank'], s=100, color='green',
alpha=0.7)
sns.regplot(x = df['CPI'], y = df['Rank'], scatter=False,
color='green')
plt.suptitle('Scatter plot entre CPI e Rank')
plt.xlabel('CPI')
plt.ylabel('Rank')
plt.show()
plt.clf()

# Scatter plot entre as variáveis Rank e V_Democracy
sns.scatterplot(x = df['Rank'], y = df['V_Democracy'], s=100,
color='blue', alpha=0.7)
sns.regplot(x = df['Rank'], y = df['V_Democracy'], scatter=False,
color='blue')
plt.suptitle('Scatter plot entre Rank e V_Democracy')
plt.xlabel('Rank')
```

```
plt.ylabel('V_Democracy')
plt.show()
plt.clf()

# Verificar a distribuição das variáveis numéricas
plt.hist(df['CPI'], color='red', alpha = 0.5, density=True, label =
'CPI')
plt.hist(df['V_Democracy'], color='green', alpha = 0.5, density=True,
label = 'V_Democracy')
plt.hist(df['Bertelsmann_TII'], color='blue', alpha = 0.5,
density=True, label = 'Bertelsmann_TII')
plt.legend(loc='upper right')
plt.show()
plt.clf()
```

Realizado ajuste na nomeclatura doas regiões para melhor visualização e entendimento dos dados

```
# Ajustar nomeclaturas das Regiões
df['Region'] = df['Region'].replace('SSA', 'Africa Sub-Sahariana')
df['Region'] = df['Region'].replace('AME', 'América')
df['Region'] = df['Region'].replace('AP', 'Ásia-Pacífico')
df['Region'] = df['Region'].replace('WE/EU', 'Europa Ocidental e
Central')
df['Region'] = df['Region'].replace('ECA', 'Europa e Ásia Central')
df['Region'] = df['Region'].replace('MENA', 'Oriente Médio e Norte da
África')

# Verificar a distribuição das variáveis categóricas
df['Region'].value_counts().plot(kind='barh', figsize=(20, 10))
plt.title('Distribuição das regiões dos países')
plt.show()
plt.clf()

# Relacionar a variável categórica Region com a média do
EIU_Country_Ratings e mostrar num gráfico de barras
df_region = df.groupby('Region')
['EIU_Country_Ratings'].mean().sort_values(ascending=False)
df_region.plot(kind='barh', figsize=(20, 10))
plt.title('Média do EIU_Country_Ratings por Região')
plt.ylabel('Região')
plt.xlabel('Média do EIU_Country_Ratings')
plt.show()
plt.clf()

# Relacionar a variável categórica Region com a média do PRS_ICRG e
mostrar num gráfico de barras
df_region = df.groupby('Region')
['PRS_ICRG'].mean().sort_values(ascending=False)
df_region.plot(kind='barh', figsize=(20, 10))
```

```
plt.title('Média do PRS_ICRG por Região')
plt.ylabel('Região')
plt.xlabel('Média do PRS_ICRG')
plt.show()
plt.clf()

# Relacionar a variável categórica Region com a média do CPI e mostrar
num gráfico de barras
df_region = df.groupby('Region')
['CPI'].mean().sort_values(ascending=False)
df_region.plot(kind='barh', figsize=(20, 10))
plt.title('Média do CPI por Região')
plt.ylabel('Região')
plt.xlabel('Média do CPI')
plt.show()
plt.clf()

# Relacionar a variável categórica Region com a média do
Global_Insights e mostrar num gráfico de barras
df_region = df.groupby('Region')
['Global_Insights'].mean().sort_values(ascending=False)
df_region.plot(kind='barh', figsize=(20, 10))
plt.title('Média do Global_Insights por Região')
plt.ylabel('Região')
plt.xlabel('Média do Global_Insights')
plt.show()
plt.clf()

# Criar bloxpot para a variável numérica V_Democracy separado por
Região
sns.boxplot(x='Region', y='V_Democracy', data=df, palette='Set1')
plt.title('Boxplot do V_Democracy por Região')
plt.show()
plt.clf()

# Criar bloxpot para a variável numérica CPI separado por Região
sns.boxplot(x='Region', y='CPI', data=df, palette='Set2')
plt.title('Boxplot do CPI por Região')
plt.show()
plt.clf()

# Criar bloxpot para a variável numérica Global_Insights
sns.boxplot(x='Region', y='Global_Insights', data=df, palette='Set3')
plt.title('Boxplot do Global_Insights por Região')
plt.show()
plt.clf()
```

Conclusões:

A análise exploratória dos dados do Índice de Percepção de Corrupção (CPI) de 2023 revelou várias informações importantes sobre a percepção de corrupção em diferentes países e regiões ao redor do mundo. Aqui estão algumas das principais conclusões e insights obtidos:

- A pontuação média do CPI foi de aproximadamente 45, com um desvio padrão de 20. Isso indica que há uma variação significativa na percepção de corrupção entre os países incluídos no dataset.
- A região com a maior média de CPI foi a Europa Ocidental e Central, seguida pela Europa e Ásia Central e pela América. A África Sub-Sahariana teve a menor média de CPI.
- A região com a maior média de Global Insights foi a Europa Ocidental e Central, seguida pela Europa e Ásia Central e pela América. A África Sub-Sahariana teve a menor média de Global Insights.
- A região com a maior média de V-Democracy foi a Europa Ocidental e Central, seguida pela Europa e Ásia Central e pela América. A África Sub-Sahariana teve a menor média de V-Democracy.
- Em termos de correlação vemos que os países com maior percepção de corrupção (CPI) tendem a uma pior avaliação no restante dos índices apresentados nos dados analisados.
- Outra análise importante é que independente da visão de democracia, temos, através da distribuição de frequência da variável V_Democracy, que nao há uma relação direta com a percepção de corrupção, ou seja, países com maior percepção de corrupção não necessariamente são menos democráticos. Um exemplo disso é a China, que está entre os 50% menos corruptos, de acordo com o CPI, mas é um país com baixa democracia (apenas um partido governa o país).

Proponha próximos passos para uma análise mais aprofundada ou para a aplicação dos dados em um modelo preditivo, se aplicável.

Para próximas análises seria interessante colocar dados de PIB Per Capita e outros dados econômicos para verificar a relação entre a percepção de corrupção e o desenvolvimento econômico dos países. Além disso, seria interessante aplicar técnicas de machine learning para prever a percepção de corrupção com base em outros índices e variáveis disponíveis no dataset. Isso poderia ajudar a identificar outras métricas que possam influenciar na percepção de corrupção e a desenvolver ações eficazes para combate a corrupção.