```python
# 1. Wczytywanie danych i wyświetlanie podstawowych informacji
import pandas as pd
df = pd.read_csv('IHME_GDP_1960_2050_Y2021M09D22.CSV')
print(df.head())
print(df.info())
print(df.describe())
```

```
   location_id location_name iso3   level  year  gdp_ppp_mean
gdp_ppp_lower  \
0            1        Global    G  Global  1960  1.748345e+13
1.601915e+13
1            1        Global    G  Global  1961  1.813537e+13
1.659537e+13
2            1        Global    G  Global  1962  1.895328e+13
1.739039e+13
3            1        Global    G  Global  1963  1.965662e+13
1.811706e+13
4            1        Global    G  Global  1964  2.100575e+13
1.935664e+13

   gdp_ppp_upper  gdp_usd_mean  gdp_usd_lower  gdp_usd_upper
0   1.911586e+13  1.296863e+13   1.266890e+13   1.334177e+13
1   1.982493e+13  1.346097e+13   1.314767e+13   1.383021e+13
2   2.061477e+13  1.406576e+13   1.376060e+13   1.443746e+13
3   2.134993e+13  1.461831e+13   1.432132e+13   1.497693e+13
4   2.276791e+13  1.552986e+13   1.523498e+13   1.587998e+13
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19838 entries, 0 to 19837
Data columns (total 11 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   location_id    19838 non-null  int64
 1   location_name  19838 non-null  object
 2   iso3           18655 non-null  object
 3   level          19838 non-null  object
 4   year           19838 non-null  int64
 5   gdp_ppp_mean   19838 non-null  float64
 6   gdp_ppp_lower  19838 non-null  float64
 7   gdp_ppp_upper  19838 non-null  float64
 8   gdp_usd_mean   19838 non-null  float64
 9   gdp_usd_lower  19838 non-null  float64
 10  gdp_usd_upper  19838 non-null  float64
dtypes: float64(6), int64(2), object(3)
memory usage: 1.7+ MB
None
        location_id          year  gdp_ppp_mean  gdp_ppp_lower
gdp_ppp_upper  \
count  19838.000000  19838.000000  1.983800e+04   1.983800e+04
1.983800e+04
mean     949.871560   2005.000000  1.334543e+12   1.235788e+12
```

```
1.444079e+12
std     5965.433243     26.268513  9.148287e+12   8.610030e+12
9.789327e+12
min        1.000000   1960.000000  1.448063e+02   6.299026e+01
2.621094e+02
25%       63.000000   1982.000000  3.678736e+03   2.639116e+03
4.829886e+03
50%      125.500000   2005.000000  1.103640e+04   8.105541e+03
1.346178e+04
75%      183.000000   2028.000000  2.949281e+04   2.308992e+04
3.562660e+04
max    44578.000000   2050.000000  1.827414e+14   1.667007e+14
2.025062e+14
```

```
       gdp_usd_mean  gdp_usd_lower  gdp_usd_upper
count  1.983800e+04   1.983800e+04   1.983800e+04
mean   8.554096e+11   8.197528e+11   8.967612e+11
std    6.286364e+12   6.041288e+12   6.585419e+12
min    1.174979e+02   8.318772e+01   1.270468e+02
25%    1.624411e+03   1.395430e+03   1.828575e+03
50%    4.863298e+03   4.279291e+03   5.465731e+03
75%    1.997525e+04   1.795003e+04   2.223434e+04
max    1.119468e+14   1.017185e+14   1.239708e+14
```

```
# 2. Obliczanie podstawowych statystyk
mean_gdp_ppp_lower = df['gdp_ppp_lower'].mean()
print('Średnia wartość gdp_ppp_lower:', mean_gdp_ppp_lower)
mean_gdp_ppp_upper = df['gdp_ppp_upper'].mean()
print('Średnia wartość gdp_ppp_upper:', mean_gdp_ppp_upper)
median_gdp_ppp_lower = df['gdp_ppp_lower'].median()
print('Mediana wartości gdp_ppp_lower:', median_gdp_ppp_lower)
median_gdp_ppp_upper = df['gdp_ppp_upper'].median()
print('Mediana wartości gdp_ppp_upper:', median_gdp_ppp_upper)
std_gdp_ppp_lower = df['gdp_ppp_lower'].std()
print('Odchylenie standardowe wartości gdp_ppp_lower:',
std_gdp_ppp_lower)
std_gdp_ppp_upper = df['gdp_ppp_upper'].std()
print('Odchylenie standardowe wartości gdp_ppp_upper:',
std_gdp_ppp_upper)
```

```
Średnia wartość gdp_ppp_lower: 1235788443809.8582
Średnia wartość gdp_ppp_upper: 1444078620521.2554
Mediana wartości gdp_ppp_lower: 8105.54082625498
Mediana wartości gdp_ppp_upper: 13461.7811907802
Odchylenie standardowe wartości gdp_ppp_lower: 8610029537548.005
Odchylenie standardowe wartości gdp_ppp_upper: 9789326655387.197
```

```
# 3. Identyfikacja i obsluga brakujących danych
missing_values = df.isnull().sum()
```

```
print('Liczba brakujących wartości w poszczególnych kolumnach:')
print(missing_values)
```

```
Liczba brakujących wartości w poszczególnych kolumnach:
location_id         0
location_name       0
iso3             1183
level               0
year                0
gdp_ppp_mean        0
gdp_ppp_lower       0
gdp_ppp_upper       0
gdp_usd_mean        0
gdp_usd_lower       0
gdp_usd_upper       0
dtype: int64
```

```
# 3. Identyfikacja i obsluga brakujących danych (c.d.)

# Uzupełnianie brakujących wartości iso3 (kod kraju) na podstawie
location_name (nazwa kraju)
df['iso3'] = df['iso3'].fillna(df['location_name'].str[:3])
missing_values = df.isnull().sum()
print('Liczba brakujących wartości w poszczególnych kolumnach po
uzupełnieniu:')
print(missing_values)
```

```
Liczba brakujących wartości w poszczególnych kolumnach po
uzupełnieniu:
location_id      0
location_name    0
iso3             0
level            0
year             0
gdp_ppp_mean     0
gdp_ppp_lower    0
gdp_ppp_upper    0
gdp_usd_mean     0
gdp_usd_lower    0
gdp_usd_upper    0
dtype: int64
```

```
# 4. Wykrywanie wartości odstających (używając metody IRQ):
Q1 = df['gdp_ppp_upper'].quantile(0.25)
Q3 = df['gdp_ppp_upper'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df['gdp_ppp_upper'] < lower_bound) |
(df['gdp_ppp_upper'] > upper_bound)]
```

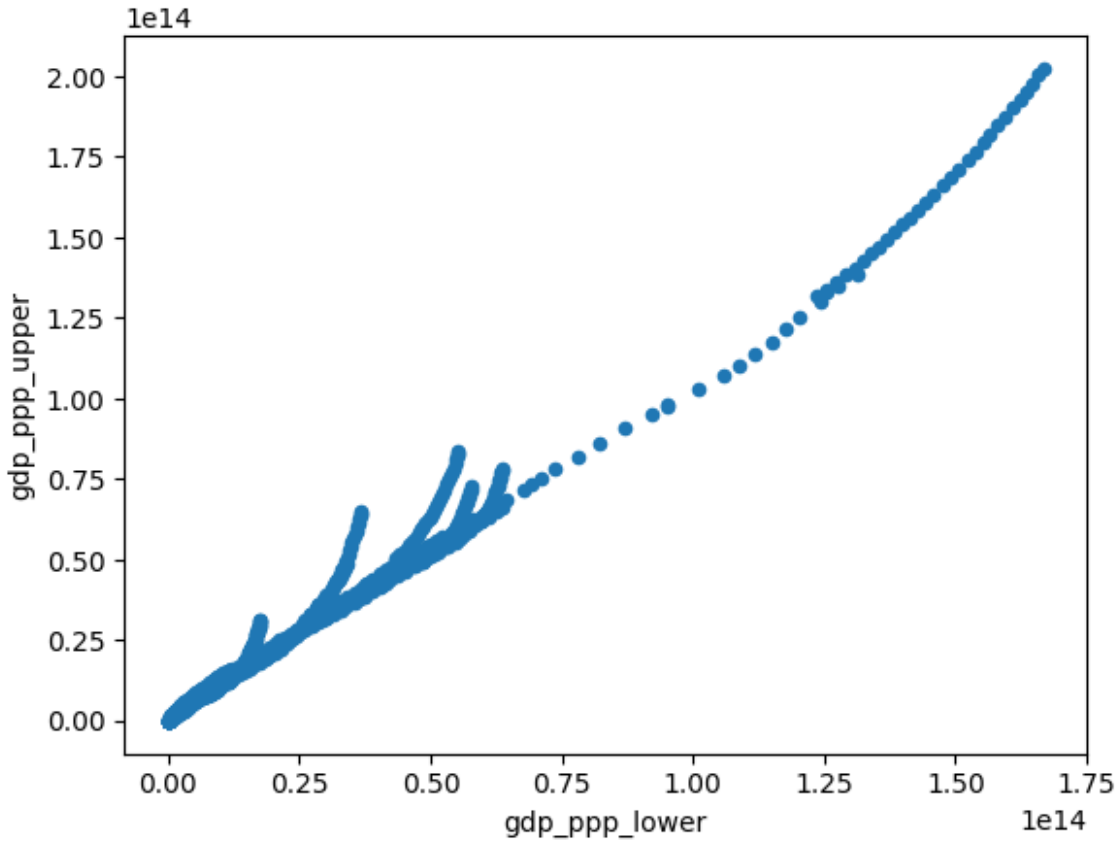
```
print('Wartości odstające w kolumnie gdp_ppp_upper:')
print(outliers)
```

```
Wartości odstające w kolumnie gdp_ppp_upper:
       location_id location_name iso3                    level
year  \
0                1        Global    G                   Global  1960

1                1        Global    G                   Global  1961

2                1        Global    G                   Global  1962

3                1        Global    G                   Global  1963

4                1        Global    G                   Global  1964

...            ...           ...  ...                      ...   ...

19833        44578    Low income  Low  World Bank Income Group  2046

19834        44578    Low income  Low  World Bank Income Group  2047

19835        44578    Low income  Low  World Bank Income Group  2048

19836        44578    Low income  Low  World Bank Income Group  2049

19837        44578    Low income  Low  World Bank Income Group  2050


       gdp_ppp_mean  gdp_ppp_lower  gdp_ppp_upper  gdp_usd_mean  \
0      1.748345e+13   1.601915e+13   1.911586e+13  1.296863e+13
1      1.813537e+13   1.659537e+13   1.982493e+13  1.346097e+13
2      1.895328e+13   1.739039e+13   2.061477e+13  1.406576e+13
3      1.965662e+13   1.811706e+13   2.134993e+13  1.461831e+13
4      2.100575e+13   1.935664e+13   2.276791e+13  1.552986e+13
...             ...            ...            ...           ...
19833  3.617310e+12   3.140835e+12   4.166469e+12  1.149318e+12
19834  3.724063e+12   3.225849e+12   4.292403e+12  1.186597e+12
19835  3.831942e+12   3.307609e+12   4.424674e+12  1.224062e+12
19836  3.941856e+12   3.398884e+12   4.560961e+12  1.262129e+12
19837  4.053883e+12   3.482933e+12   4.713596e+12  1.300764e+12

       gdp_usd_lower  gdp_usd_upper
0       1.266890e+13   1.334177e+13
1       1.314767e+13   1.383021e+13
2       1.376060e+13   1.443746e+13
3       1.432132e+13   1.497693e+13
4       1.523498e+13   1.587998e+13
...              ...            ...
19833   1.031500e+12   1.271992e+12
19834   1.061313e+12   1.318836e+12
```

```
19835   1.092874e+12   1.365610e+12
19836   1.122895e+12   1.413991e+12
19837   1.151548e+12   1.457362e+12

[1901 rows x 11 columns]
```

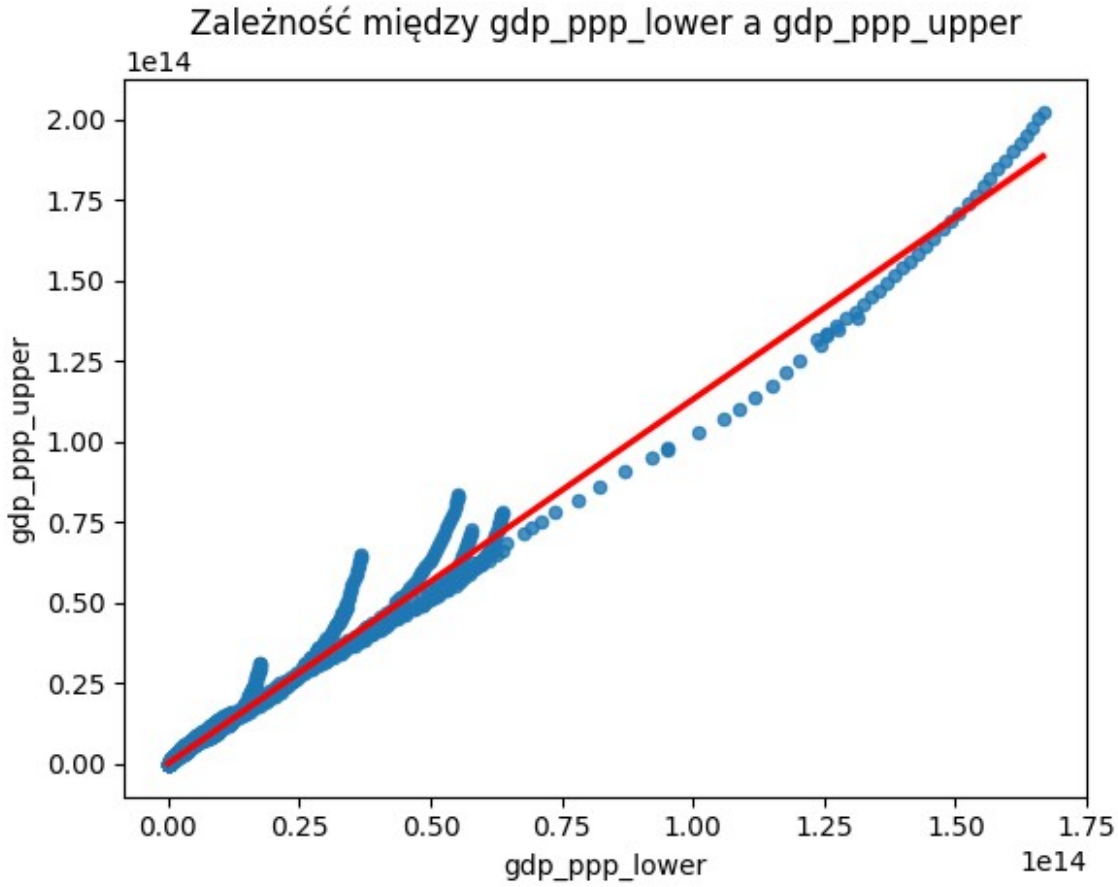
```python
# 5. Analiza zależności między kolumnami
numeric_df = df.select_dtypes(include=['number']) # wybierz tylko kolumny numeryczne
correlation = numeric_df.corr()
# print('Macierz korelacji:')
# print(correlation)
numeric_df.plot.scatter(x='gdp_ppp_lower', y='gdp_ppp_upper')
```

```
<Axes: xlabel='gdp_ppp_lower', ylabel='gdp_ppp_upper'>
```

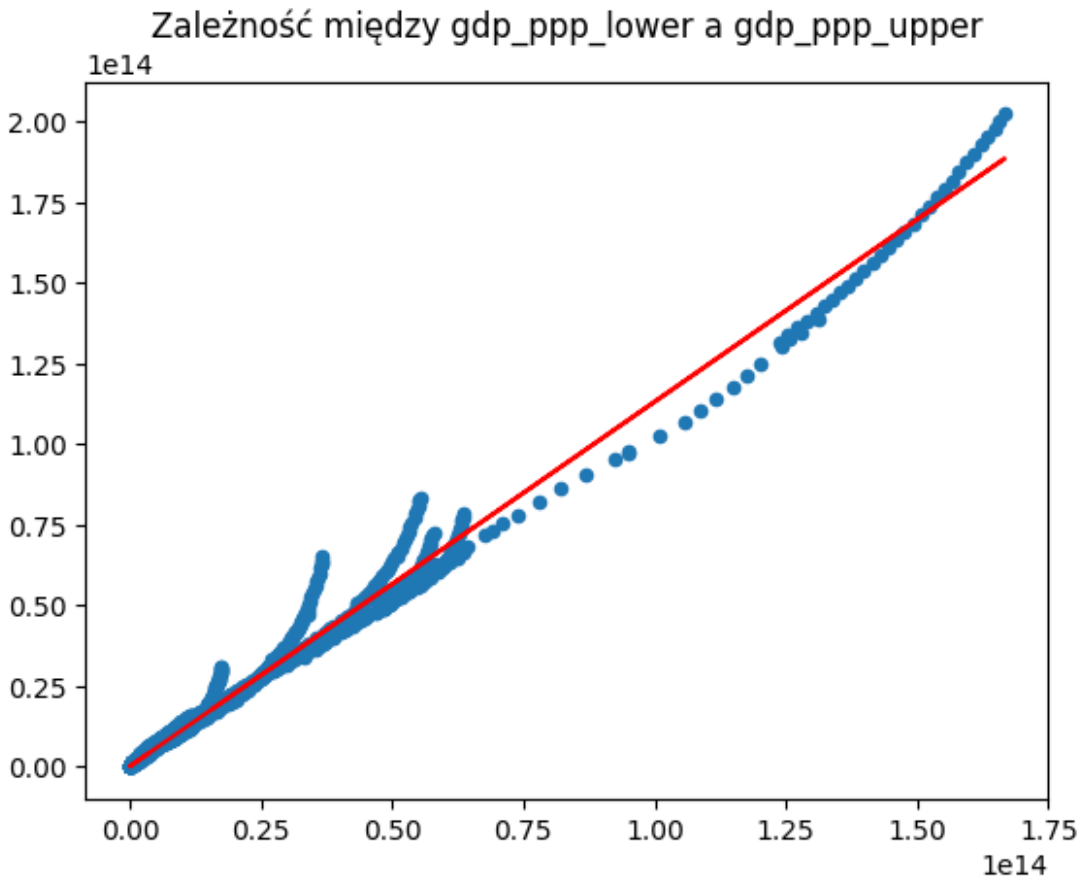



```python
# 5. Analiza zależności między kolumnami (z linią regresji)

import seaborn as sns
import matplotlib.pyplot as plt

numeric_df = df.select_dtypes(include=['number']) # wybierz tylko kolumny numeryczne
```

```
sns.regplot(x='gdp_ppp_lower', y='gdp_ppp_upper', data=numeric_df,
scatter_kws={'s': 20}, line_kws={'color': 'red'})
plt.title('Zależność między gdp_ppp_lower a gdp_ppp_upper')
plt.show()
```




```
# 5. Analiza zależności między kolumnami (alternatywnie)
import numpy as np
x = numeric_df['gdp_ppp_lower']
y = numeric_df['gdp_ppp_upper']
slope, intercept = np.polyfit(x, y, 1) # Oblicz współczynniki regresji
(liniowa regresja)
plt.scatter(x, y, s=20)
plt.plot(x, slope * x + intercept, color='red')
plt.title('Zależność między gdp_ppp_lower a gdp_ppp_upper')
plt.show()
```




```
# 6. Przekształcanie danych
df['gdp_ppp_diff'] = df['gdp_ppp_upper'] - df['gdp_ppp_lower'] #
Dodanie nowej kolumny

mean_diff = df.groupby('location_name')['gdp_ppp_diff'].mean() #
Grupowanie wg nazwy kraju i obliczanie średniej wartości gdp_ppp_diff:
print('Średnia różnica między gdp_ppp_upper a gdp_ppp_lower dla
poszczególnych krajów:')
print(mean_diff)

# Sortowanie po kolumnie year:
df = df.sort_values(by='year', ascending=True)
print(df.head())
```

```
Średnia różnica między gdp_ppp_upper a gdp_ppp_lower dla
poszczególnych krajów:
location_name
Afghanistan                                1413.412350
Albania                                    3310.023453
Algeria                                    4607.599093
American Samoa                             3479.864825
Andorra                                   11830.056188
```

```
                                          ...
Venezuela (Bolivarian Republic of)     8096.820512
Viet Nam                               4157.516836
Yemen                                  2963.275206
Zambia                                 1817.071337
Zimbabwe                               1829.976392
Name: gdp_ppp_diff, Length: 216, dtype: float64
       location_id                      location_name iso3    level
year  \
13832          171  Democratic Republic of the Congo  COD  Country
1960
6097            79                           Finland  FIN  Country
1960
10465          132                            Panama  PAN  Country
1960
4459            58                           Estonia  EST  Country
1960
15015          185                            Rwanda  RWA  Country
1960

       gdp_ppp_mean  gdp_ppp_lower  gdp_ppp_upper  gdp_usd_mean  \
13832   2529.408870    1497.917422    4030.682914   1192.389928
6097   13070.254728   10573.992294   15651.094527  13228.211834
10465   5315.615945    3772.621329    7078.008722   2825.271221
4459   10175.164360    7461.704732   15956.346985   5571.152548
15015   1047.273739     758.179810    1471.468483    285.107907

       gdp_usd_lower  gdp_usd_upper  gdp_ppp_diff
13832     986.677359    1402.576098   2532.765492
6097    12068.494529   14689.939429   5077.102233
10465    2570.171286    3120.499562   3305.387392
4459     5340.938458    5856.943423   8494.642253
15015     221.465258     334.067442    713.288674
```