衡量一个现代化湾区的发展水平可以从经济、科技和人口方面进行分析。下采用pearson系数来衡量各因素与GDP间的相关性

import pandas as pd  
import os  
import scipy.stats as stats

data\_dir = os.path.join(os.getcwd(), 'data')  
economy\_factor\_dir = os.path.join(data\_dir, '经济')  
population\_factor\_dir = os.path.join(data\_dir, '人口')  
SciTech\_factor\_dir = os.path.join(data\_dir, '科技')

GDP\_file = pd.read\_excel(os.path.join(data\_dir, 'GDP.xlsx'))  
GDP = GDP\_file['大湾区汇总'][:-1]  
GDP.index = list(range(1999,2024))  
print(GDP)

1999 19980.230000  
2000 22389.990000  
2001 23321.990000  
2002 24519.570000  
2003 26184.840000  
2004 29508.700000  
2005 33522.030000  
2006 38124.660000  
2007 44011.270000  
2008 48973.530000  
2009 50794.090000  
2010 58051.950000  
2011 66188.760000  
2012 71760.760000  
2013 78973.260000  
2014 84829.650000  
2015 90121.260000  
2016 96706.290000  
2017 105593.180000  
2018 113250.280000  
2019 120108.670000  
2020 118633.020563  
2021 132073.579634  
2022 134745.950000  
2023 143922.770000  
Name: 大湾区汇总, dtype: float64

下对经济相关因素进行相关性分析，首先进行数据处理，将所有的NA值进行线性差值处理。遇到数据缺失导致数据长度不同时采用截断处理。

with open(os.path.join(os.getcwd(), "Economic\_Factors.csv"), 'a') as f:  
 f.write("经济相关因素,Pearson系数,p值\n")  
for files in os.listdir(economy\_factor\_dir):  
 file\_path = os.path.join(economy\_factor\_dir, files)  
 factor\_file = pd.read\_excel(file\_path)  
 factor = factor\_file['大湾区汇总']  
 factor = factor.reset\_index(drop=True)  
 factor.index += 1999  
 # 填充NA值  
 factor = factor.interpolate(method='linear')  
 # 截断处理  
 GDP\_e = GDP  
 if factor.shape[0] < GDP\_e.shape[0]:  
 GDP\_e = GDP\_e[:factor.shape[0]]  
 else:  
 factor = factor[:GDP\_e.shape[0]]  
   
 corr, p\_val = stats.pearsonr(factor, GDP\_e)  
 with open(os.path.join(os.getcwd(), "Economic\_Factors.csv"), 'a') as f:  
 f.write(f"{os.path.splitext(files)[0]},{corr},{p\_val}\n")

下对人口相关因素进行相关性分析，处理方法同上。

with open(os.path.join(os.getcwd(), "Population\_Factors.csv"), 'a') as f:  
 f.write("人口相关因素,Pearson系数,p值\n")  
for files in os.listdir(population\_factor\_dir):  
 file\_path = os.path.join(population\_factor\_dir, files)  
 factor\_file = pd.read\_excel(file\_path)  
 factor = factor\_file['大湾区汇总']  
 factor = factor.reset\_index(drop=True)  
 factor.index += 1999  
 # 填充NA值  
 factor = factor.interpolate(method='linear')  
 # 截断处理  
 GDP\_p = GDP  
 if factor.shape[0] < GDP\_p.shape[0]:  
 GDP\_p = GDP\_p[:factor.shape[0]]  
 else:  
 factor = factor[:GDP\_p.shape[0]]  
   
 corr, p\_val = stats.pearsonr(factor, GDP\_p)  
 with open(os.path.join(os.getcwd(), "Population\_Factors.csv"), 'a') as f:  
 f.write(f"{os.path.splitext(files)[0]},{corr},{p\_val}\n")

下对科技相关因素进行相关性分析，由于2015年前的科技数据缺失严重，故仅使用2015年后的数据。其他处理方法同上

with open(os.path.join(os.getcwd(), "SciTech\_Factors.csv"), 'a') as f:  
 f.write("科技相关因素,Pearson系数,p值\n")  
for files in os.listdir(SciTech\_factor\_dir):  
 file\_path = os.path.join(SciTech\_factor\_dir, files)  
 factor\_file = pd.read\_excel(file\_path)  
 factor = factor\_file['大湾区汇总']  
 factor = factor.reset\_index(drop=True)  
 factor.index += 2015  
 # 填充NA值  
 factor = factor.interpolate(method='linear')  
 factor = factor[0:6+1] # 2015~2021  
 # 截断处理  
 GDP\_st = GDP  
 GDP\_st = GDP\_st[16:23]  
  
 corr, p\_val = stats.pearsonr(factor, GDP\_st)  
 with open(os.path.join(os.getcwd(), "SciTech\_Factors.csv"), 'a') as f:  
 f.write(f"{os.path.splitext(files)[0]},{corr},{p\_val}\n")