广东金融学院实验报告

课程名称：数据分析与数据挖掘

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验编号  及实验名称 | 实验一 数据分析 | | | 系 别 | 互联网金融 |
| 姓 名 | 林旋华 | 学号（完整学号） | 181543306 | 班 级 | 18计科3班 |
| 实验地点 | 实验楼102 | 实验日期 | 2021.04.19 | 实验时数 | 2 |
| 指导老师 | 侯昉 | 同组其他成员 | 无 | 成 绩 |  |
| 一、实验目的及要求  1、巩固Python程序编制，Numpy、Pandas和Matplotlib包的使用。  2、掌握基本的数据获取、整理和分析流程及其操作。 | | | | | |
| 二、实验环境及相关情况（包含使用软件、实验设备、主要仪器及材料等）  1．使用软件：Python开发工具  2．实验设备：个人计算机 | | | | | |
| 三、实验内容及步骤（包含简要的实验步骤流程）  1.使用Pandas\_datareader获取任意两支股票近三个月的交易数据。做出收盘价的变动图像。  2.使用Pandas\_datareader获取世界银行数据库中美国（USA）、瑞典（SWE）、瑞士（CHE）三个国家近20年的NY.GDP.PCAP.KD数据，作图分析。  3.对于泰坦尼克的数据集，试分析幸存与否与独立登船的相关性（alone数据列）。 | | | | | |
| 四、实验结果（包括程序或图表、结论陈述、数据记录及分析等，可附页）  1.程序如下：  #导入pandas\_datareader/datetime/seaborn/matplotlib库  import pandas\_datareader.data as webdata  import datetime  import seaborn as sns  import matplotlib.pyplot as plt  #处理Seaborn图表内嵌中文字体问题  plt.rcParams['font.sans-serif']=['SimHei']  plt.rcParams['axes.unicode\_minus'] = False  sns.set\_style({'font.sans-serif':['simhei','Arial']})  #用datatime生成日期数据  start\_time = datetime.datetime(2021,1,30)  end\_time = datetime.datetime(2021,3,30)  #通过Yahoo财经，查询股票信息  stock\_code1 = input('美股如GOOG \n港股，如腾讯：0700.hk \n国内股票代码后面加.ss或者.sz\n你要查询的股票代码是:')  stock\_info1 = webdata.get\_data\_yahoo(stock\_code1,start\_time,end\_time)  stock\_code2 = input('另一只要查询的股票代码是:')  stock\_info2 = webdata.get\_data\_yahoo(stock\_code2,start\_time,end\_time)  #绘制股票价格波动图  print('股票收盘价对比：')  sns.set()  plt.figure(figsize=(10,4))  plt.plot(stock\_info1['Close'],'r-',label=stock\_code1)  plt.plot(stock\_info2['Close'],'k--',label=stock\_code2)  plt.legend()  plt.savefig('股票收盘.png')  plt.show()    注意：导入模块时需要由import pandas.io.data as web更改为import pandas\_datareader.data as web,需要先下载pandas\_datareader包。   1. 程序如下：   #导入wb用于查询世界银行数据  import pandas\_datareader.wb as worldbank  import matplotlib.pyplot as plt  import pandas as pd  import seaborn as sns  #处理Seaborn图表内嵌中文字体问题  sns.set()  sns.set\_style({'font.sans-serif':['simhei','Arial']})  plt.rcParams['font.sans-serif']=['SimHei']  plt.rcParams['axes.unicode\_minus'] = False  start\_year = 1999  end\_year = 2019  #查询数据库中所有国家信息  #countries = worldbank.get\_countries()  #读取美国（USA）、瑞典（SWE）、瑞士（CHE）的NY.GDP.PCAP.KD数据  dat = worldbank.download(indicator="NY.GDP.PCAP.KD", country=["USA", "SWE","CHE"], start=start\_year, end=end\_year)  dat2draw=dat.unstack(level=0)  #dat2draw.plot()  plt.figure(figsize=(10,4))  plt.plot(dat2draw.iloc[:,0],'r-',label="SWE")  plt.plot(dat2draw.iloc[:,1],'b-\*',label="CHE")  plt.plot(dat2draw.iloc[:,2],'g--',label="USA")  plt.title("人均GDP对比（美元计价）",fontsize=20)  plt.legend()     1. 程序如下：   import matplotlib.pyplot as plt  from matplotlib import pyplot as plt  import seaborn as sns  import pandas as pd  #处理Seaborn图表内嵌中文字体问题  plt.rcParams['font.sans-serif']=['SimHei']  plt.rcParams['axes.unicode\_minus'] = False  sns.set\_style('whitegrid',{'font.sans-serif':['simhei','Arial']})  #载入泰坦尼克数据库，显示乘客信息的头5行  PasTit = pd.read\_csv("titanic.csv")  print('前5行数据：')  print(PasTit.head())  #统计幸存者是否独立登船  survived=PasTit.groupby(['alone','survived']).size().unstack()  survived['sum']=survived[0]+survived[1]  survived['生还率']=survived[1]/survived['sum']  print()  print('乘客按照是否独立登船：')  print('False：不是独立登船；True：独立登船')  print(survived)  #幸存与否与独立登船相关性分析(相关系数)  X=PasTit['survived']  Y=PasTit['alone']  print(X.corr(Y))  #生成相关性热力图  plt.subplots(figsize=(9, 5)) # 设置画面大小  sns.heatmap(survived, annot=True, vmax=1, square=True, cmap="Blues")  plt.savefig('热力图.png')  plt.show() | | | | | |
| 五、实验总结（包括心得体会、问题回答及实验改进意见，可附页）  通过本次实验，我巩固了Python的基础知识，Numpy、Pandas、seaborn和Matplotlib包的使用。  对使用Pandas\_datareader获取基本的数据有了进一步认识，并掌握使用webdata.get\_data\_yahoo的方法从yahoo财经获取财经信息；学会使用Matplotlib、seaborn做出收盘价的变动图像和相关系数热力图；发现了一些要注意的点和疑惑的地方，对dataframe生成、索引切片和matplotlib画图。  在本次实验中，由于浏览器没有访问 https://github.com/mwaskom/seaborn-data 网站的权限，导致titanic.csv数据集下载不到，幸好之前有把下载的数据集存档，但目前还未找到解决方案，留存上课询问老师。同时发现自己的基础还不是很扎实，课后会加强巩固基础知识，争取更好！ | | | | | |
| 六、教师评语  1、完成所有规定的实验内容，实验步骤正确，结果正确；  2、完成绝大部分规定的实验内容，实验步骤正确，结果正确；  3、完成大部分规定的实验内容，实验步骤正确，结果正确；  4、基本完成规定的实验内容，实验步骤基本正确，所完成的结果基本正确；  5、未能很好地完成规定的实验内容或实验步骤不正确或结果不正确。  评定等级：  签名： | | | | | |