宋琪语-10235501465-“数据科学与工程导论 - 06"

2024年9月30日 18:03

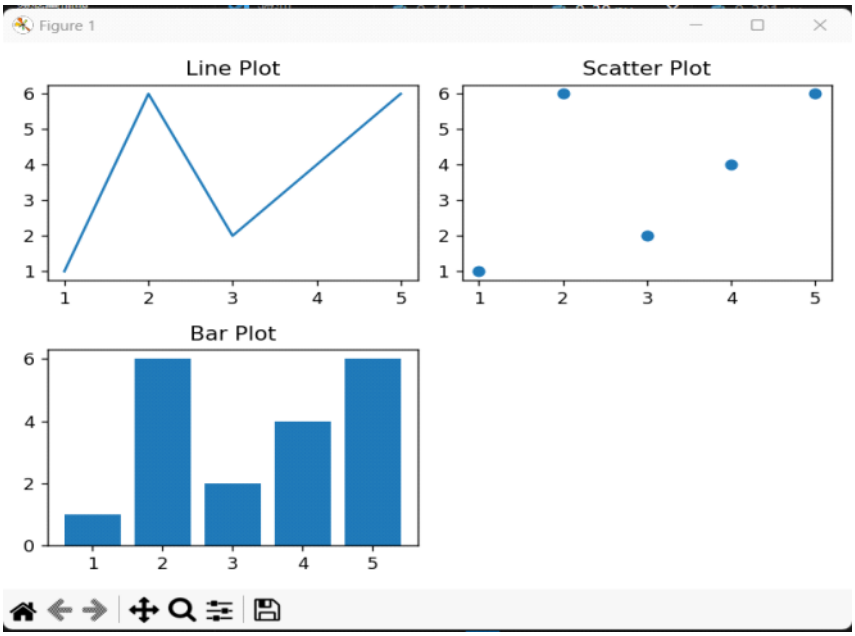
1. 数据采集，数据存储，数据管理，数据计算，数据分析，数据展示
2. 数据采集是指从真实世界对象中获得原始数据的过程。最常用的三种数据采集方法是：传感器、日志文件和爬虫。
3. 数据管理是利用计算机硬件和软件技术对数据进行有效的采集，存储，处理和应用的过程。异：大数据管理的本质是在大数据基础设施层之上，向用户提供更加方便、高效、友好的人机交互界面。随着大数据处理技术的发展，大数据的终端用户将越来越不用关心底层的基础设施、计算、存储、网络等细节，只需关心数据本身，以及如何开展大数据应用的创新与服务。同：都是数据管理，都涉及到了对数据有效的采集，存储，处理和应用。
4. 数据的计算模式大致分为批量计算模式、流式计算模式、交互式计算模式和图计算模式四类。
5. 数据分析处理来自对某一兴趣现象的观察、测量或实验的信息，其目的是从与主题相关的数据中提取尽可能多的信息，主要目标包括推测或解释数据并确定如何使用数据、检查数据是否合法、给决策提供合理建议、诊断或推断错误原因以及预测未来将要发生的事情。
数据分析的方法：描述性分析，预测性分析，规律性分析。
数据分析的模型：统计模型，机器学习模型，时间序列模型，聚类分析
6. 数据可视化的原因是我们利用视觉获取的信息量，远远比别的感官要多得多；数据可视化能够在小空间中展示大规模数据；数据可视化能够帮助我们对数据有更加全面的认识；受人类大脑记忆能力的限制，可视化之后能更好的记忆。

7.

```
import matplotlib.pyplot as plt
import numpy as np
x = [1, 2, 3, 4, 5]
y = [1, 6, 2, 4, 6]
plt.figure()
plt.subplot(2, 2, 1)
plt.plot(x, y)
plt.title('Line Plot')

plt.subplot(2, 2, 2)
plt.scatter(x, y)
plt.title('Scatter Plot')

plt.subplot(2, 2, 3)
plt.bar(x, y)
plt.title('Bar Plot')
plt.tight_layout()
plt.show()
```



8.

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
x = [1, 2, 3, 4, 5]
y = [1, 6, 2, 4, 6]
plt.figure()
plt.subplot(2, 2, 1)
sns.set_theme(style="darkgrid")
sns.barplot(x=x, y=y)
plt.xlabel('柱状图')
plt.subplot(2, 2, 2)
with sns.axes_style("ticks"):
    plt.pie(y, labels=x, autopct='%1.1f%%')
plt.xlabel('饼图')
plt.subplot(2, 2, 3)
sns.pointplot(x=[1, 2, 3], y=[4, 5, 6])
plt.xlabel('折线图')
sns.set_style({"font.sans-serif": ['simhei', 'Droid Sans Fallback']})
plt.tight_layout()
plt.show()
```