



**《Python玩转大数据》课程大作业**

**实**

**施**

**报**

**告**

|  |  |
| --- | --- |
| **项目名称：** | **足球五大联赛球员伤病数据分析及预测模型** |
| **学生学号：** | **2022216590** |
| **学生姓名：** | **吴秉康** |
| **所属班级：** | **精细化工22-2** |

**二O二五年五月**

**《足球五大联赛球员伤病数据分析及预测模型》实施报告**

**摘要：** 本项目围绕足球五大联赛（英超、西甲、德甲、意甲、法甲）球员的伤病数据展开分析与预测。通过编写Python爬虫，从transfermarkt.com获取了球员的基本信息和伤病历史数据，并构建了统一的数据结构。项目以球员的比赛特征和伤病历史为输入，通过可视化技术直观展示伤病类型、频率和时序分布。进一步地，结合序列建模的思路，尝试设计机器学习模型对伤病趋势进行预测。项目实现了从数据获取、清洗、分析到可视化与预测的完整流程。尽管由于数据维度限制和技术条件影响，模型仍有一定局限性，但为后续更专业的伤病分析与体育数据挖掘提供了良好基础，并展现了将AI技术应用于竞技体育健康管理的广阔前景。

**关键字：**足球；五大联赛；球员伤病；数据分析；机器学习；可视化；序列建模；网络爬虫

**一、项目简介**

本项目旨在基于五大联赛（英超、西甲、德甲、意甲、法甲）球员数据，建立伤病分析与预测模型。通过编写Python爬虫，从 transfermarkt.com 获取包括球员基本信息、比赛数据与伤病记录在内的多维数据。项目对数据进行清洗、结构化建模与可视化展示，深入分析不同类型伤病的分布与影响。进一步地，基于球员属性和伤病历史构建时间序列特征，尝试引入机器学习方法对未来伤病风险进行预测。项目不仅提升了对足球运动中伤病问题的理解，也为运动科学与数据分析的结合提供了实用范例。

项目代码：https://github.com/ericdonno/python-bigdata-course-proj.git

**二、项目环境及资源的配置**

操作系统：Windows 11

Python环境：Anaconda 2024.10

开发工具：VS code、Jupyter Notebook

数据来源：www.transfermarkt.com

**三、项目实施过程**

具体技术细节在Notebook中由markdown写出。

**1、数据集的准备**

从获取原始数据，再到数据的初步清洗、整理、规范化、标准化……

json 文件：

player\_dict.json (url) ->injury\_dict (url) -> injury\_data\_dict (data) -> feature\_data\_dict (data)

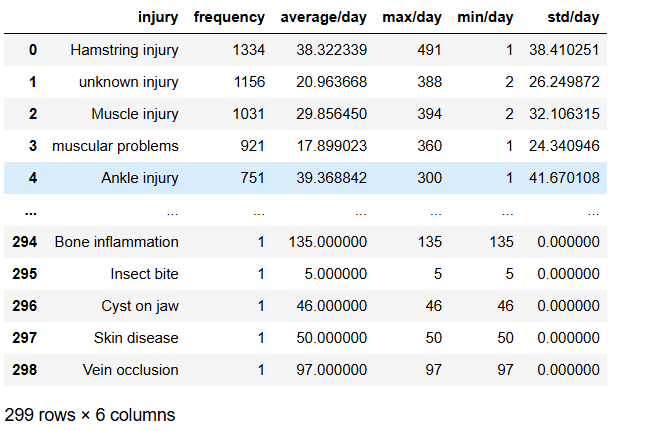
经过了两次非常漫长的爬取过程。每次request设置了5~10s的间隔，大约会耗时2\*7.5s\*2591 = 10.8h。最后得到的两个字典为很高质量的数据。

其他的json文件起辅助作用。

**2、数据统计分析处理**

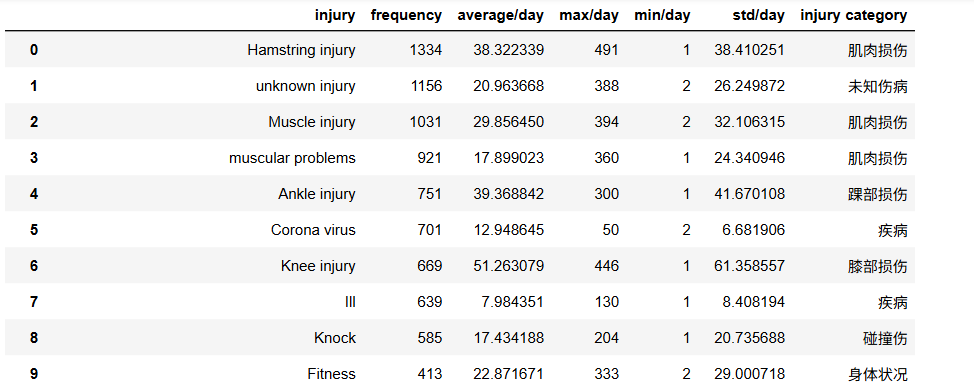
使用相关的规则或分析指标体系，对数据进行统计分析。

**伤病信息的统计分析**

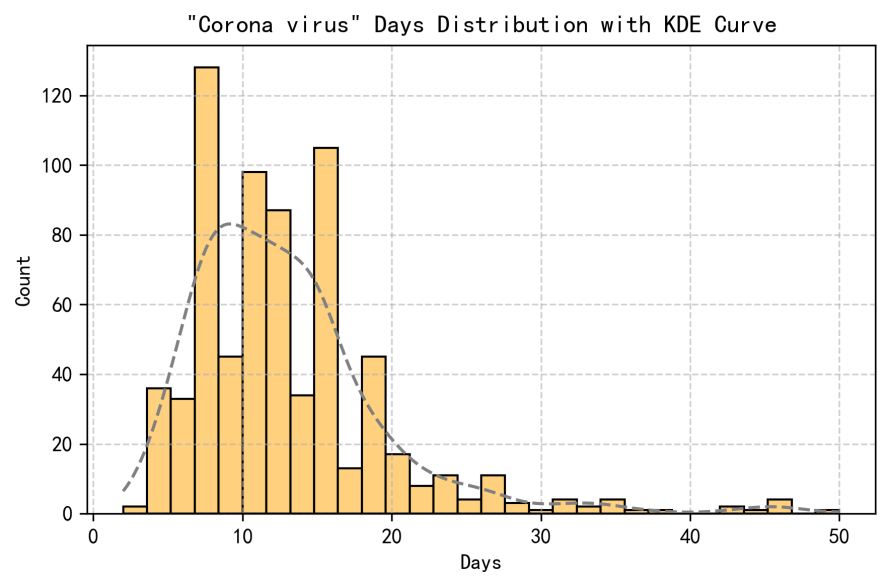


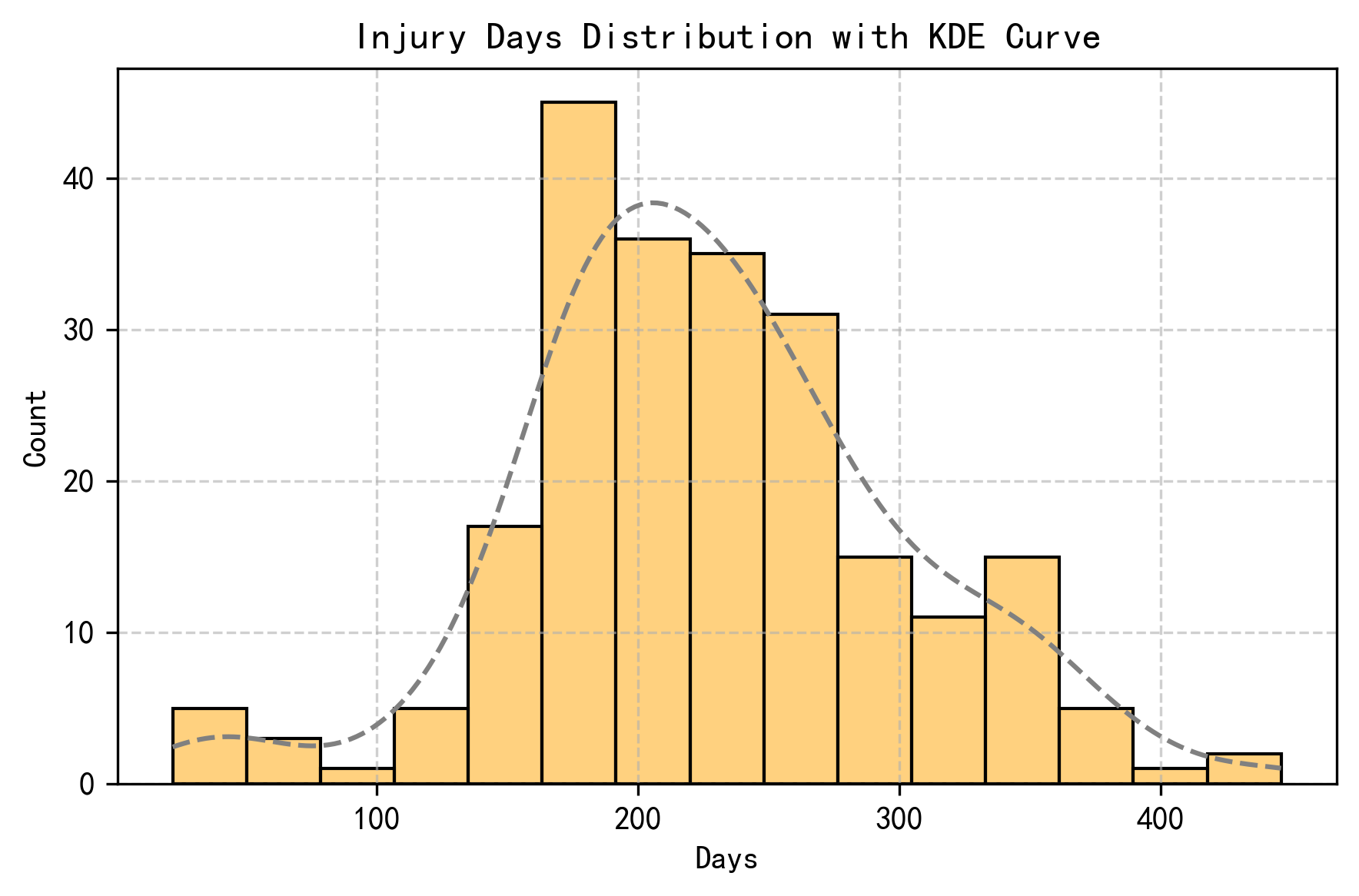
对不同伤病类别进行了频数，平均值，最大值，最小值，标准差的统计。

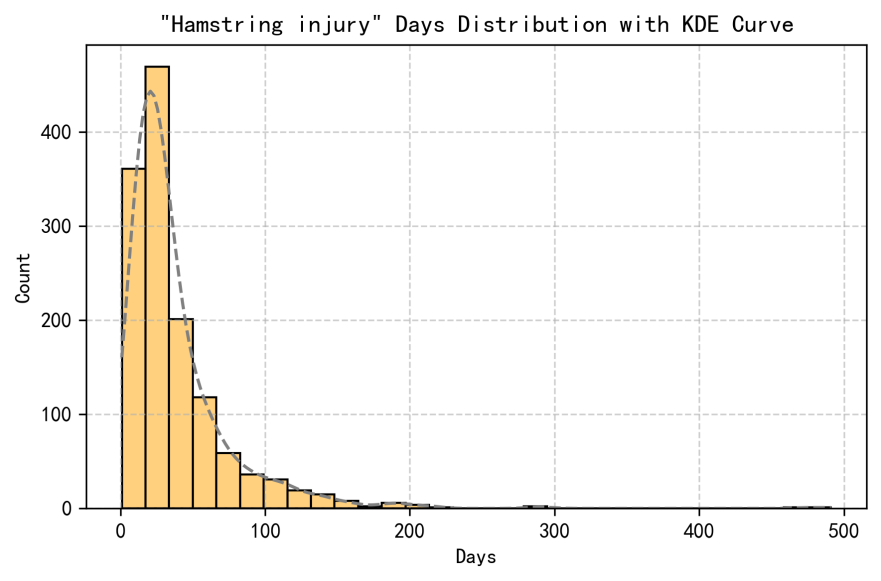
将伤病类型分了大类进行分析。

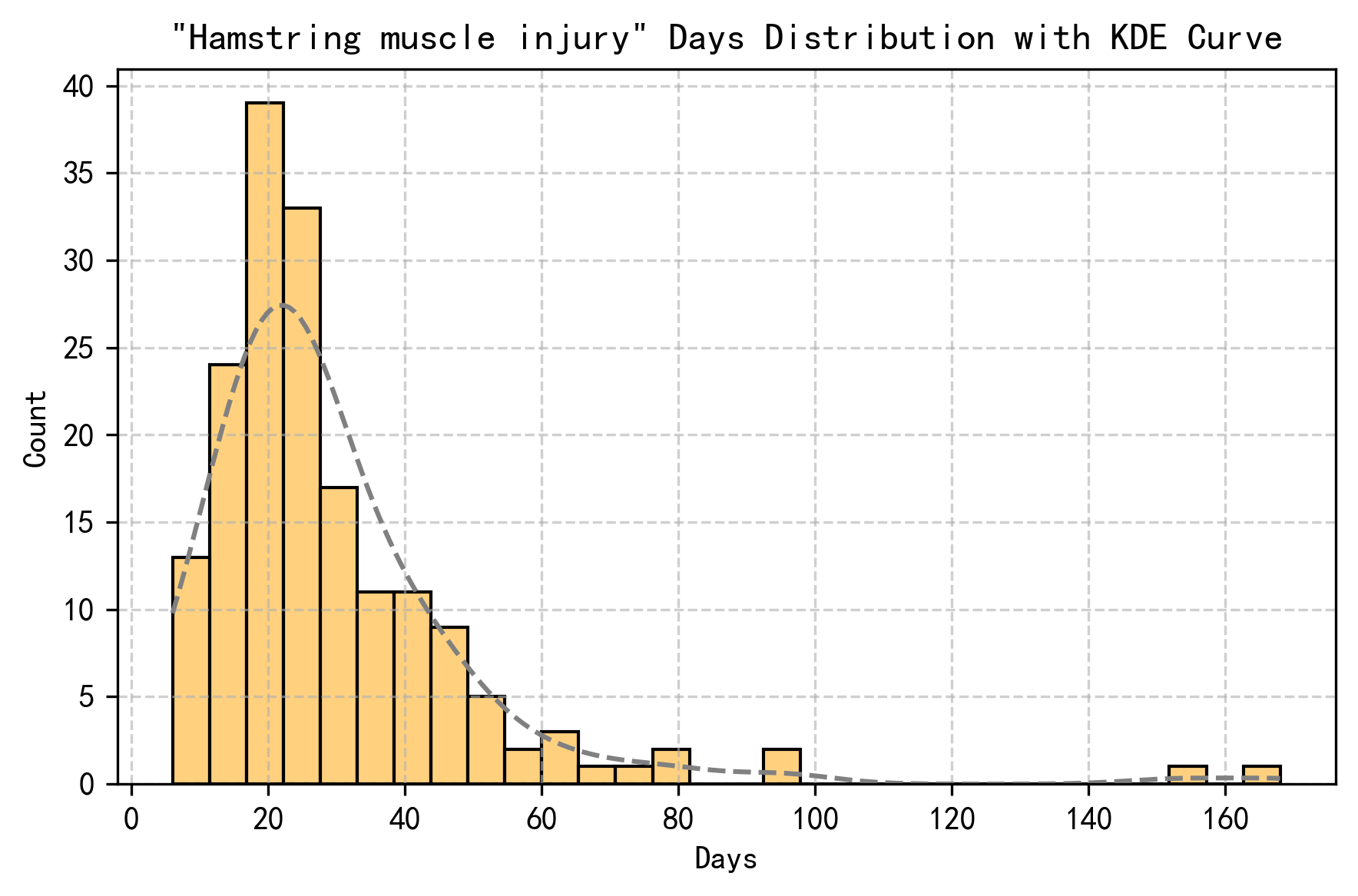


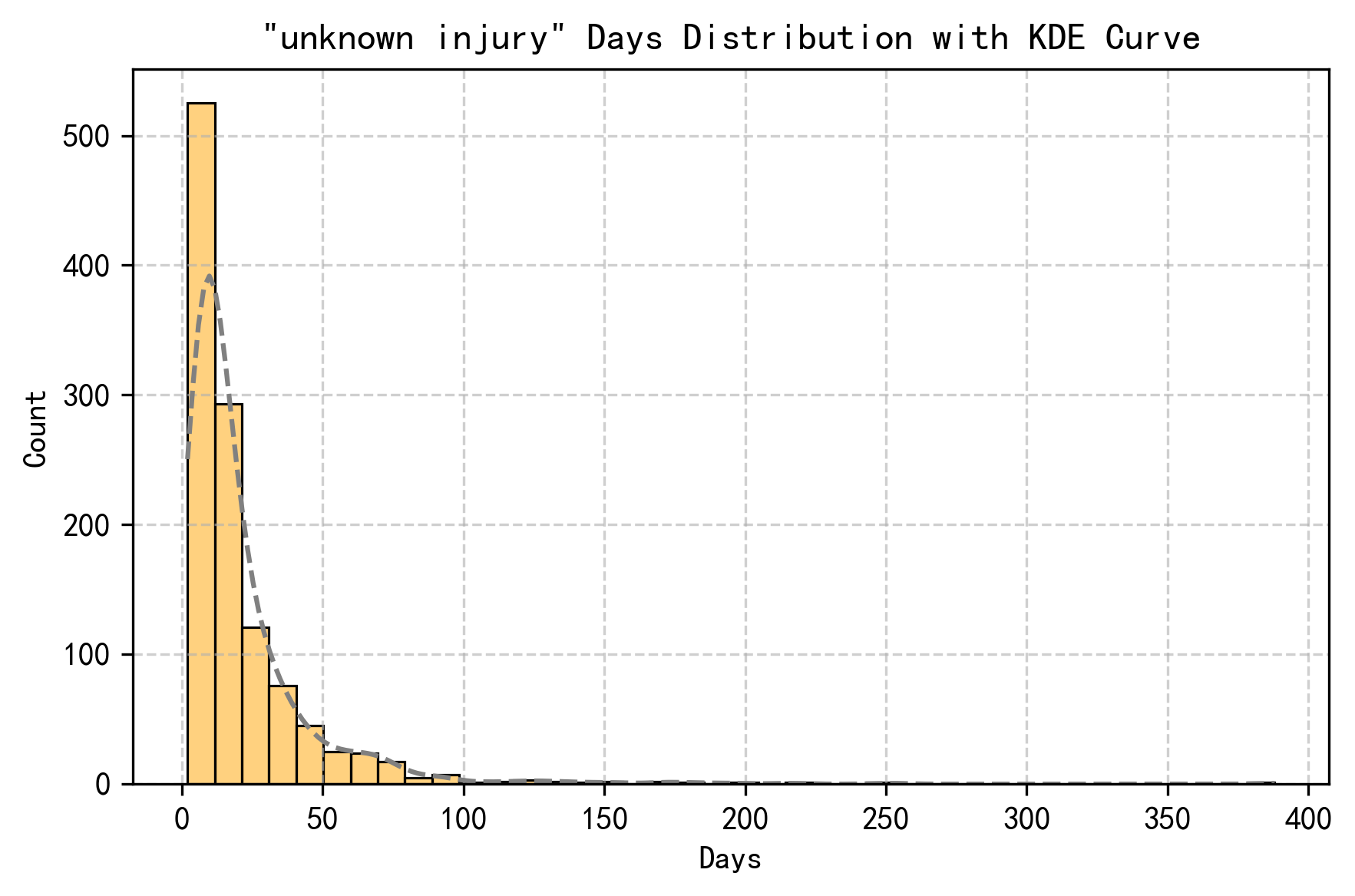
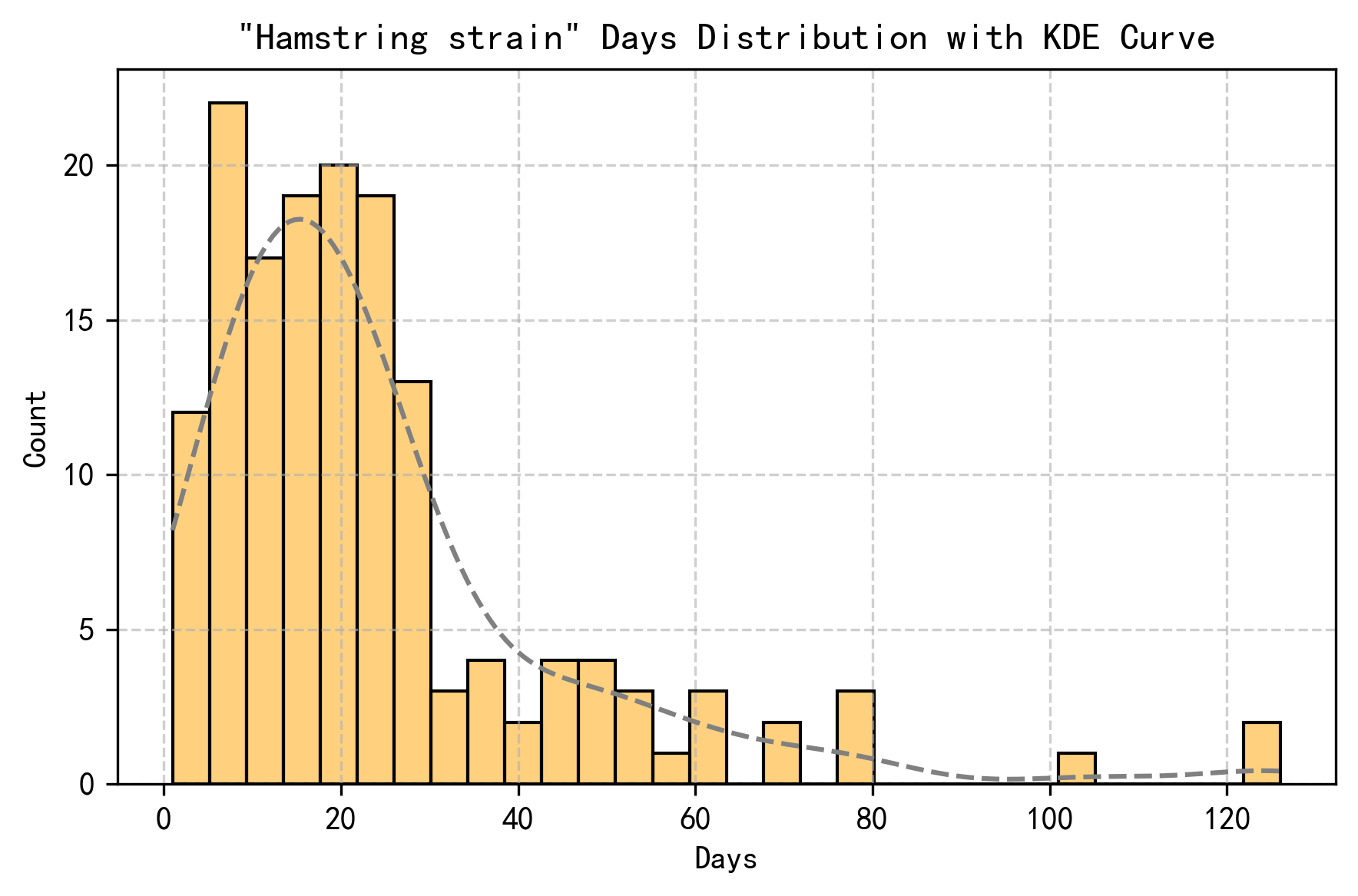
**3、数据处理结果可视化**

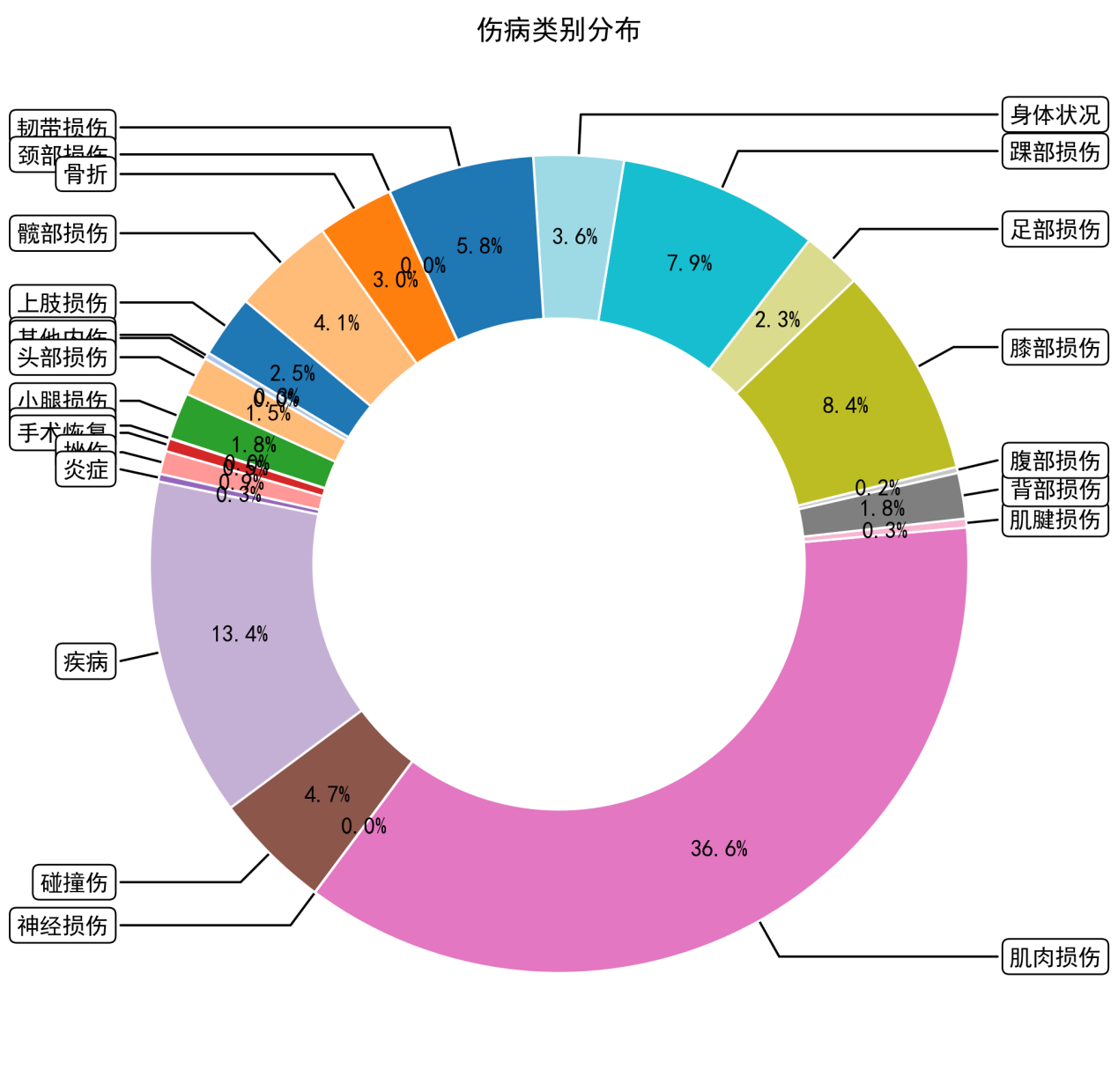
****

****

****

****

****



**4、数据处理分析的结论或对数据可视化的剖析**

伤病的时长分布类似泊松分布，但其中重大伤病可能会导致赛季报销的，如“十字韧带撕裂”类似于正态分布。



困扰运动员伤病的类别主要在于肌肉损伤，属于是高频率，较低时长的伤病。球员最不希望的还是韧带、脚踝和膝盖伤势，这些伤势会严重影响球员的出勤。

**四、未来展望**

通过此项目的研究，对未来相关问题或学习思考的感受。该项目能够让我们对足球运动员的伤病，通过数据分析，有着更清晰的认知。伤病虽然残酷，但仍是竞技运动的一部分。我们需要理性地看待它，科学地预防它。

足球是及其复杂的系统，但是一般来说对其的评价体系是纯经验主义的，缺乏科学性的。

该项目虽说有一定量的数据和技术支撑，但由于各种限制，其分析与模型仍是偏经验的。对伤病专业的预测需要更精细的数据和模型。球员的身体运动模式、肌肉骨骼结构，以及心理因素、场外因素都是我们无法获取同时也是关键的数据。所以说该项目存在局限性，从而导致其一定的非专业性。

通过该项目的完成，我们有了可靠的数据来源与较为成熟的处理流程，可以更轻松得完成足球相关的其他类型的数据分析与可视化，如球队战术风格分析、球员身价及状态分析、球员与球队适配度分析等等。

该项目也可以进一步扩大规模，获取其他联赛的球员数据。甚至经过简单修改，进行其他体育项目运动员的伤病分析及预测。另外还可以设计GUI方便输入输出，甚至封装为产品。