实践报告：

成都租房数据分析与可视化

学号：2213189

姓名：李懋

日期：2024.10.11

南开大学 软件学院

目 录

目录

[成都租房数据分析与可视化 1](#_Toc382443166)

[1 实验介绍 2](#_Toc303735491)

[2 实验目的 3](#_Toc313768318)

[3 实验流程 4](#_Toc1431767241)

[4 实验结果或结论 5](#_Toc1157754852)

[12](#_Toc36472097)

# 实验介绍

根据课程网站上给出的例子,爬取成都市的租房价格信息,保存到 CSV 文件中.再将已有的数据进行数据分析和可视化,进而对成都市的租房情况有一个初步的了解.

# 实验目的

- 利用爬虫技术从链家网站上爬取成都市租房价格情况,将数据爬下储存到CSV文件中.后续再对数据读取,预处理清洗,可视化分析等操作.

- 进而可以直观的对成都租房价格,区域分布等等数据有个更深入的了解,帮助人们根据自己的条件,选择租房的最优解.

# 实验流程

## 爬取成都市房价信息

- 使用Selenium库来爬取链家网站上成都市的租房信息，并将结果保存到一个CSV文件中。

## 对已有的租房信息进行数据分析与可视化

1. 利用pandas库读取文件
2. 进行数据预处理,过滤数据
3. 使用 Matplotlib Seaborn 库进行数据可视化操作

# 实验结果或结论

## 爬取成都市租房价格信息

### 导入相关的库

### 

### 初始化 CSV 文件

### 

### 4.1.3 初始化Selenium WebDriver

### 

### 4.1.4 爬取多页数据

### 

### 4.1.5 查找租房信息,并在每段爬取后暂停一段时间

### 

### 4.1.6 爬取完成,关闭 CSV 文件和 WebDriver

### 

## 可视化分析

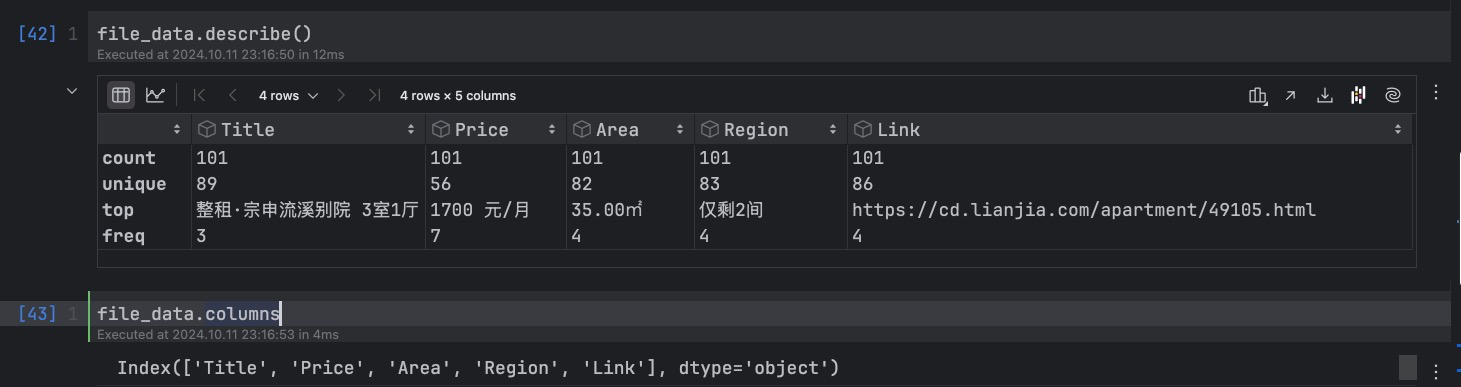
### 读取数据

- 使用pandas读取 CSV 文件中的数据

### 

- 查看数据文件中信息





### 数据预处理

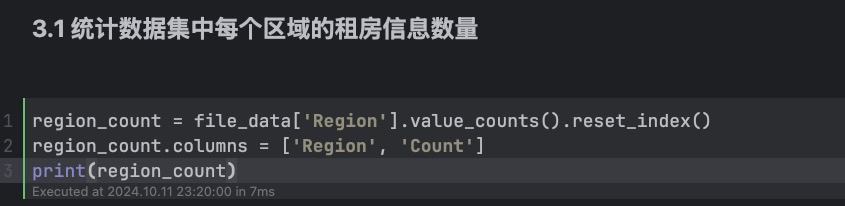
- 对数据集中的“价格”和“面积”列进行清洗和处理，以便后续分析。

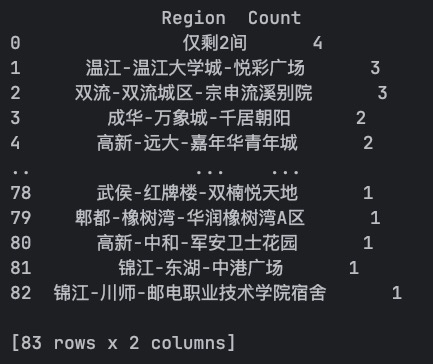
### 

### 

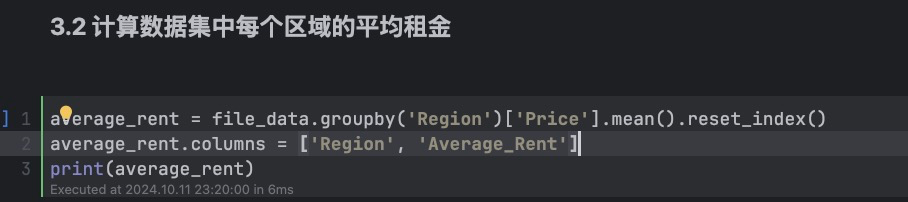
### 数据分析

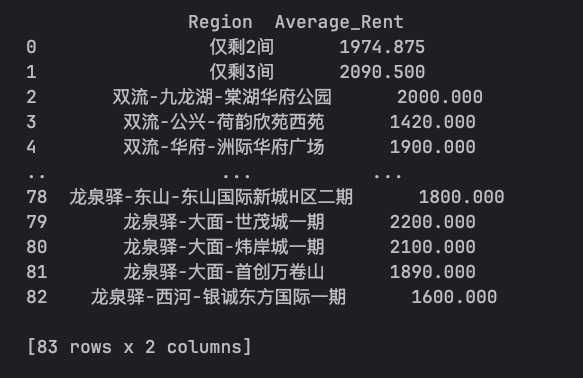
1. 统计数据集中每个区域的租房信息数量





1. 计算数据集中每个区域的平均租金



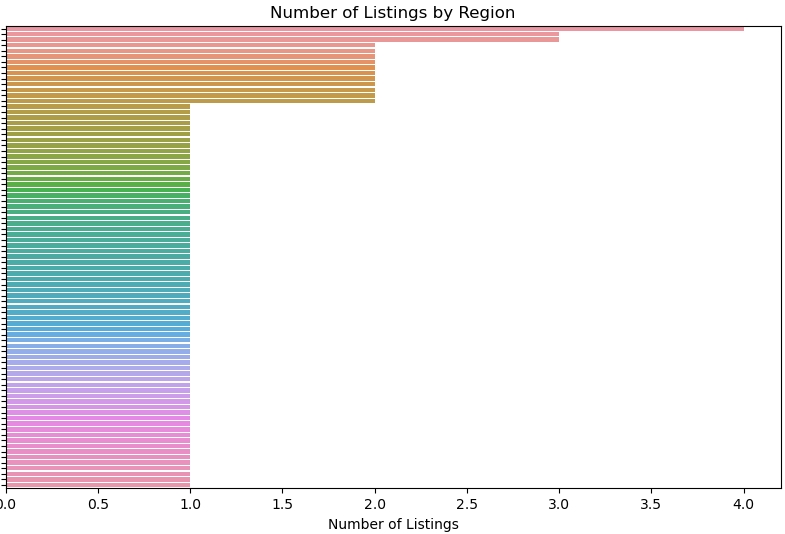


1. 计算数据集中每个区域的平均租金，将结果存储在一个新的DataFrame中，并打印出来



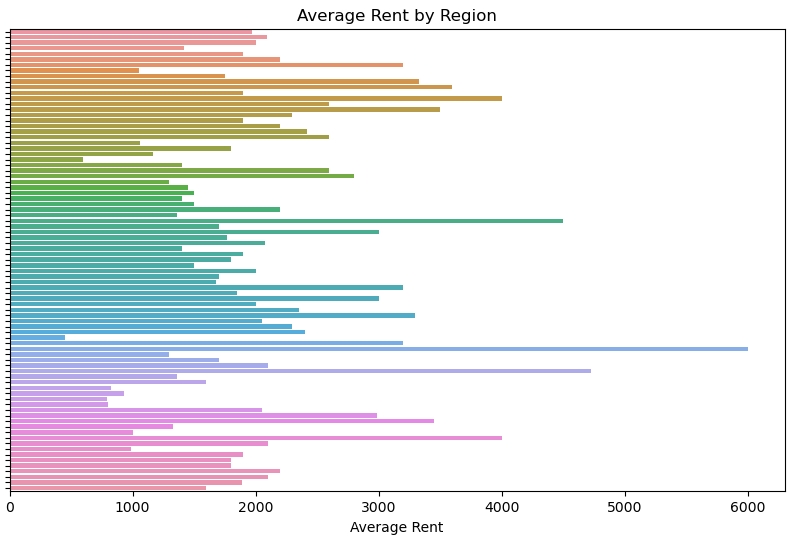
1. 图表展示不同区域的房源数量





1. 条形图展示不同区域的平均租金





## 实验总结

### 步骤总结

1. 爬取数据:使用selenium库直接从链家平台上爬取信息,存储在CSV文件中
2. 数据分析与可视化
   1. 读取数据:使用pandas库
   2. 数据预处理,清洗
   3. 使用 matplotlib seaborn 库实现数据可视化分析

### 遇到的困难

1. 在爬取数据的过程中,如果在短时间内多次访问服务器的话,会被拦截,导致爬虫失败
2. matplotlib seaborn 库在macOS系统上的中文字体显示出了问题,由于近期才换了macOS,还在配置环境和习惯中,所以这方面花费了一些时间
3. 实验过程中python解释器出现了问题,搜了很多教程也没得到解决,最后通过重启 Pycharm 问题自然就解决了.
4. word 书写实验报告还是感觉不是很高效,可能是我使用word不太熟练,感觉使用 Markdown 或 LaTeX 更加高效,排版更加美观.
5. Hw1所有代码和相关文件全部上传到 Github 仓库https://github.com/KaiHaverz/dataScience\_learningRecord

