实践3 Python数据分析与识别实践2

1、实验目的

(1) 掌握Python智能数据分析的流程

(2) 掌握数据缺失值的处理办法、掌握数据特征抽取方法

(3) 掌握机器学习模型在实际数据背景下的应用

(4) 学会自己分析给定的数据

2、实验平台

操作系统：Windows或Ubuntu

软件：Python 2.7或Python 3.6搭配Anaconda+Jupyter或Pycharm等

3、实验内容和要求

（一）基于基站定位数据的商圈分析

阅读所给的文档，了解基于基站定位数据的商圈分析过程，了解数据的背景知识，以及通过层次聚类计算，在给定的数据集上，分别完成如下步骤（背景讲解文档中存在部分代码错误，请使用所给的代码PDF版本），并将结果截图：

(1) 通过离差标准化进行基站商圈数据的归一化：

(2) 通过图示的方式画出不同计算方式下，层次聚类的谱系图：

(3) 根据谱系图获得聚类数量为3较为合适，编写代码实现商圈的层次聚类分析：

(4)（选做）感兴趣的同学，可以尝试sklearn.cluster中更多具有代表性的聚类方法，如K均值、谱聚类等方法，对比与层次聚类方法的结果。

（二）爱彼迎民宿数据的分析与建模

Airbnb爱彼迎的愿景是创造一个“家在四方”的世界，致力于帮助旅行者们在世界各地旅行时找到归属感。近日，Airbnb爱彼迎宣布，其房东已经在全球超过600万套房源中开门迎客。数据集来自 Airbnb 于 2019 年 4 月 17 日公开的北京地区数据。数据均来源于 Airbnb 网站的公开信息，不包含任何个人隐私数据。

数据集包含三部分：

* listings: 数据为北京短租房源基础信息，包括房源、房东、位置、类型、价格、评论数量和可租是时间等等。明细版中包含更多房源相关细节。
* calendar: 数据为短租房源时间表信息，包括房源、时间、是否可租、租金和可租天数等等。
* reviews: 数据为短租房源的评论信息。汇总版中仅包括房源 listing\_id和评论日期，用来时间序列和数据可视化分析。明细版还包括评论相关的内容和作者信息。

阅读所给的文档，了解数据综合分析的技巧和方法。在给定的数据集上，分别完成如下步骤并将结果截图：

（1）针对calendar数据子集，进行日期与价格的分析：

（2）针对listings数据子集，进行租房信息特征分析，并选择合适的特征：

（3）（选做）感兴趣的同学，通过网络搜索的方式，分析reviews数据子集：

思考：给定具有一定背景的数据集后，如何进行数据清洗？缺失值如何处理？异常值如何处理（如机票价格为0属于异常值）？如何进行数据可视化展示？如何选择合适的特征进行后续的机器学习模型构建？ 如何选择合适的模型进行分析如何进行结果展示？

# 4.实验报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 《人工智能技术应用实践》上机实践3报告 | | | | |
| 题目： |  | 姓名 |  | 日期 |
| 实验环境： | | | | |
| 实验内容与完成情况： | | | | |
| 出现的问题： | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | | | |