**基于假设检验的餐馆数据相关分析**

**摘要**

本文基于美国一家餐馆在244天内的数据，包括总金额，小费，性别，是否抽烟，日期以及用餐时间。餐馆负责人为了更好的调研用户用餐反馈，旨在给出数据，进行相关分析以及餐馆未来的优化规划。

**针对问题一**：我们使用了python中Pandas, Seaborn 等数据库进行了数据可视化，包括但不限于展示了总金额和小费的描述性统计矩阵，散点图，Q\_Q图等等。在此基础上，我们使用了**JB—正态分布检验**来分别检验了总金额和小费各自是否满足正态分布的规律，依据假设检验，给出原假设和备择假设，在经过分析后，我们以95%的概率拒绝了原假设，两者各自均不满足正态分布，所以考虑使用**斯皮尔曼相关系数**来对两者进行相关分析，最后我们创新的通过斯皮尔曼系数与**热力度图**结合，发现了总金额和小费有显著的正相关关系。

**针对问题二**：我们使用了Seaborn库中的**箱型图**以及**一元线性回归模型**来进行了两者的显著性分析，由于两者之间是独立的，所以我们结合**独立样本t检验**，最终发现了男性顾客在部分离散值下展现出给出更多小费的倾向，但是总体而言，男性顾客和女性顾客的消费没有什么显著性差异，所以我们认为男性顾客会比女性顾客更加慷慨。

**针对问题三**：我们将数据集中的时间进行两两分组，确定这些组别各自的相关分析结果，使用独立t检验，创新的使用在箱型图中引入中位数，平均值，离群值等等指标结合显著性标记，来总结得到了周四的总金额最高这一结论。

**针对问题四**：我们将是否抽烟和男女性别组别进行组合得到四组新数据，基于该新数据进行相关分析，结合箱型图得出结论:抽烟的男性顾客比不抽烟的男性顾客更加慷慨,抽烟的女性顾客比不抽烟的女性顾客更加慷慨。

**针对问题五**：我们将时间Dinner 和 Lunch 时间段的小费数据，从当前列进行分割，成为独立的数据，经过相关分析，箱型图绘制得到结论：晚餐时段的小费要高于午餐时段。

**关键词**：**JB—正态分布检验，斯皮尔曼相关系数，热力度图，箱型图，一元线性回归模型，独立样本t检验**

1. **问题重述**

这家美国餐馆在过去244天里，详细记录了顾客的用餐数据，包括总消费金额、小费、性别、吸烟习惯，以及用餐的具体日期和时间。这些数据的收集，旨在帮助餐馆管理层更全面地了解顾客的用餐偏好和行为模式。通过对这些数据的深入分析，餐馆希望能够识别出服务中的不足之处，并据此制定出针对性的改进措施。此外，这些信息还将用于预测顾客需求，优化菜单设计，提升顾客满意度，并最终推动餐馆业务的持续增长和市场竞争力的提升。

1. **问题分析**

**2.1分析小费和总金额之间的关系**

针对小费和总金额之间的关系，首先引入小费和总金额的平均值，最大值，最小值，中位数，众数，偏度系数，峰度系数，标准差等描述统计量，并通过数据可视化分析，通过绘制散点图来观测小费与总金额之间存在的关系。然后引入Q\_Q图和构造JB统计量来分析数据是否符合正态分布，接着引入斯皮尔曼相关系数进行相关性分析。

**2.2分析性别对慷慨度的影响**

针对性别对慷慨度的影响，首先对数据进行分类，男顾客一类，女顾客一类。然后通过数据可视化分析，通过绘制箱型图分析性别对慷慨度的影响，并构造t统计量进行独立样本t检验来分析男顾客和女顾客的小费金额是否存在显著性差异，并把显著性检验结果标到箱型图上。

**2.3分析日期和小费的关系**

针对日期与小费的关系，首先对数据进行分类，周四一类，周五一类，周六一类，周日一类。然后通过数据可视化分析，通过绘制箱型图分析日期对小费的影响，并构造t统计量进行独立样本t检验来分析周四，周五，周六，周日四者之间的小费金额是否存在显著性差异，并把显著性检验结果标到箱型图上。

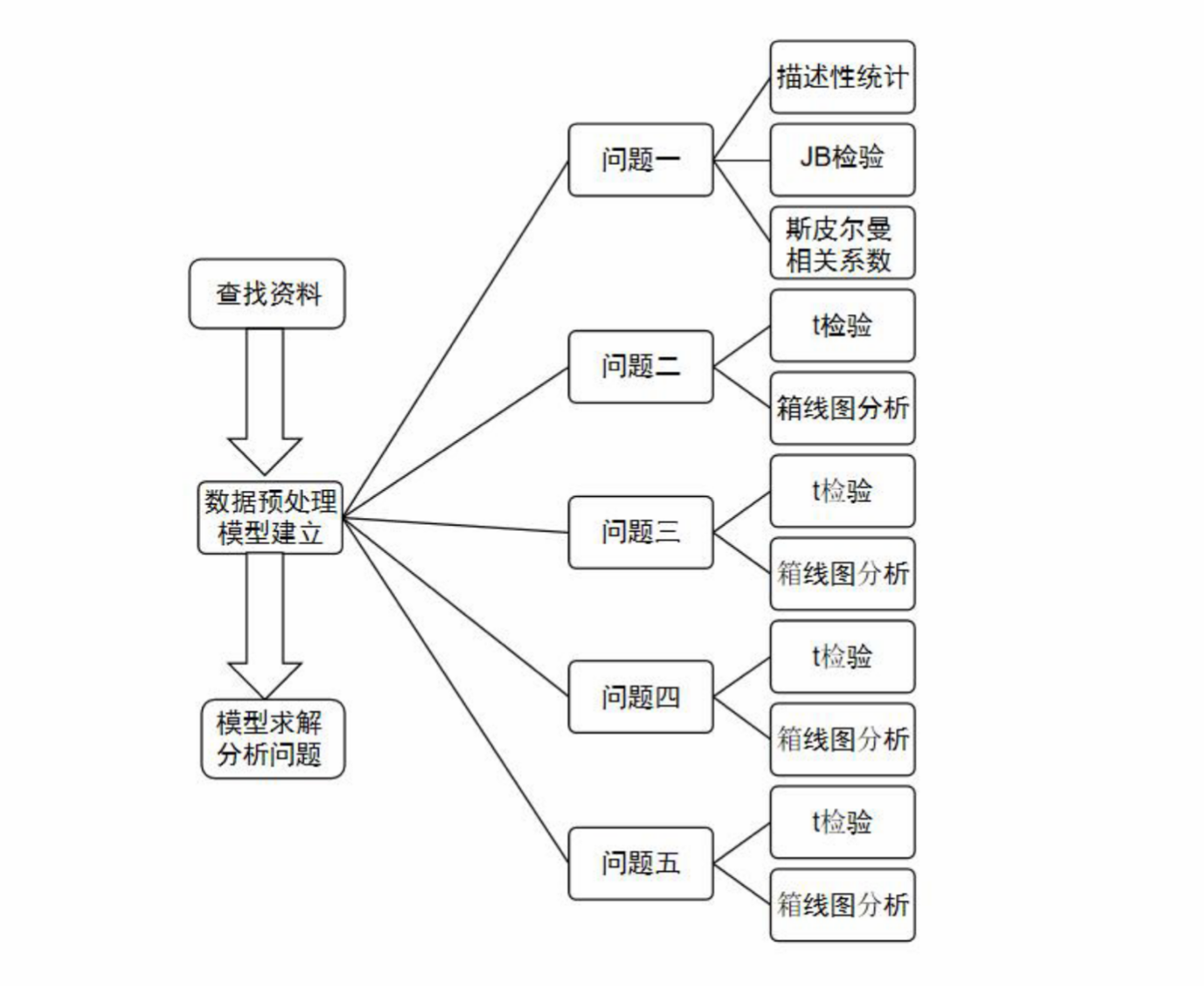
**2.4分析性别加上抽烟组合对慷慨度的影响**

针对性别加上抽烟组合对慷慨度的影响，首先对数据进行分类，抽烟的顾客一类，不抽烟的女顾客一类，抽烟的男顾客一类，不抽烟的男顾客一类。然后通过数据可视化分析，通过绘制箱型图分析性别加上抽烟组合对小费的影响，并构造t统计量进行独立样本t检验来分析抽烟的女顾客，不抽烟的女顾客，抽烟的男顾客，不抽烟的男顾客四者之间的小费金额是否存在显著性差异，并把显著性检验结果标到箱型图上。

**2.5分析就餐时间段对小费的影响**

针对就餐时间对小费的影响，首先对数据进行分类，午餐一类，晚餐一类。然后通过数据可视化分析，通过绘制箱型图分析就餐时间对小费金额的影响，并构造t统计量进行独立样本t检验来分析午餐时间和晚餐时间的小费金额是否存在显著性差异，并把显著性检验结果标到箱型图上。

文章的总体思路如图1所示:

图1总体思路图

1. **模型假设**

* 假设每条记录(每笔交易)是独立的，即每位顾客消费情况不会影响其他顾客的消费情况。
* 假设回头客去的频率在研究期间保持稳定，不存在特殊情况。
* 假设每位参与统计的顾客在餐厅消费完后都会给小费。
* 假设慷慨度的唯一衡量标准就是小费。
* 假设顾客的抽烟习惯以及性别在研究期间是稳定的。

1. **符号说明**

|  |  |
| --- | --- |
| 符号 | 意义 |
| ***S*** | 偏度系数 |
| ***K*** | 峰度系数 |
|  | 显著性水平 |
|  | 标准差 |
|  | 斯皮尔曼相关系数 |
|  | 等级差异 |

1. **模型的建立和求解**

**5.2问题一：分析总花费和小费的关系**

**5.1.1 数据预处理**

根据题目要求给出描述性统计信息，本文引入小费和总金额的平均值，最大值，最小值，中位数，众数，偏度系数，峰度系数，标准差等描述统计量

1. **偏度系数**

***偏度系数( Skewness)***用于衡量数据分布的偏斜程度。*Sk* = 0 表示数据近似对称分布,*Sk* > 0 表示数据呈右偏分布，*Sk* < 0表示数据呈左偏分布。偏度系数的绝对值越大，数据分布的偏斜程度越明显。其计算公式如下:

其中是偏度系数，*E(X)*是均值，是三阶中心距。

1. **峰度系数**

***峰度系数( Kurtosis )***用于衡量数据分布的峰度程度。*Ku* = 0表示数据分布为正态分布，*Ku* < 0 表示数据分布的峰度较小，数据分布更分散，*Ku* > 0表示数据分布的峰度较大数据更集中。峰度系数的绝对值越大，数据分布的峰度程度越明显。其计算公式如下:

其中是峰度系数，*E*(*X* )是均值，是四阶中心距。

1. **数据求解**

**利用python实现的描述性统计了如下表所示**

**表1**

|  |  |  |
| --- | --- | --- |
| 指标 | 总金额 | 小费 |
| 平均值 | 19.78 | 2.99 |
| 最大值 | 50.81 | 10.00 |
| 最小值 | 3.07 | 1.00 |
| 标准差 | 8.90 | 1.38 |
| 标准误差 | 0.57 | 0.09 |
| 偏度系数 | 1.13 | 1.46 |
| 峰度系数 | 1.22 | 3.64 |
| 中位数 | 17.79 | 2.90 |
| 众数 | 13.42 | 2.00 |

通过计算结果可知，总金额的标准差比小费的标准差大，说明总金额的波动幅度大,而小费的波动幅度小。总金额和小费的偏度都大于 0，说明数据分布相对均值向右倾斜.总金额和小费的峰度都大于 0，说明数据分布相对于正态分布较为尖峰。

**5.1.2 检测是否符合正态分布**

首先我们先绘制出总金额和小费的散点图，然后绘制出它们各自的Q\_Q图,再利用JB检验检验正态分布情况

1. 总金额和小费散点图

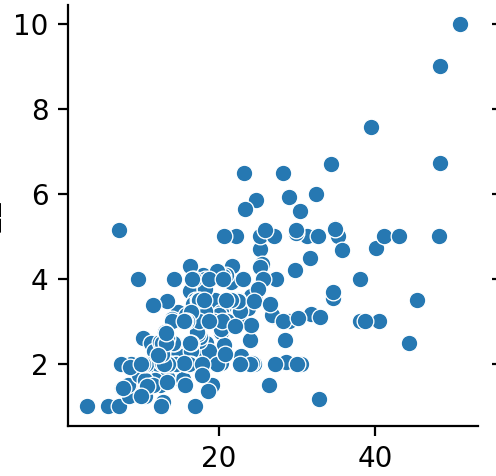


图2总金额和小费散点图

通过图2总金额和小费散点图的关系，猜想总金额和小费之间可能存在某种线性关系

1. Q\_Q图

图表, 折线图

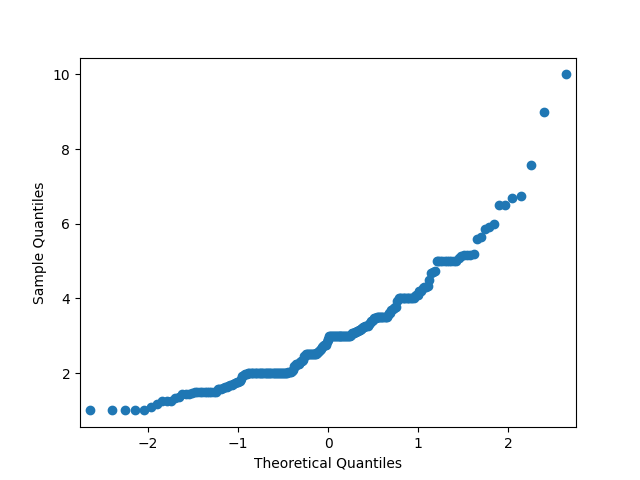
描述已自动生成

图3总金额Q\_Q图 图4小费Q\_Q图

1. JB检验

对于一个随机变量{},假设其偏度为S，峰度为K,我们也可以构造JB统计量：

可以证明，如果{}服从正态分布，那么在大样本的情况下JB统计量服从自由度为2的卡方分布，正态分布的偏度为0，峰度为3，假设检验步骤如下：  
***Step1***:提出假设

原假设:小费和总金额服从正态分布

备择假设:小费和总金额不服从正态分布

设定置信水平为95%(即显著性水平< 0.05)

***Step2***:计算p值

本文使用python进行JB检验，结果如表2所示：

表2

|  |  |  |
| --- | --- | --- |
|  | 总金额 | 小费 |
| P | 6.044824589572461e-15 | 2.841482511551161e-47 |
| JB | 65.4791478574108 | 214.35434688727574 |

由表2知，两者的p值小于显著性水平0.05，故而以95%的概率拒绝原假设，可以认为总金额和小费均不服从正态分布

**5.1.3 利用斯皮尔曼相关系数分析总金额与小费的相关性**

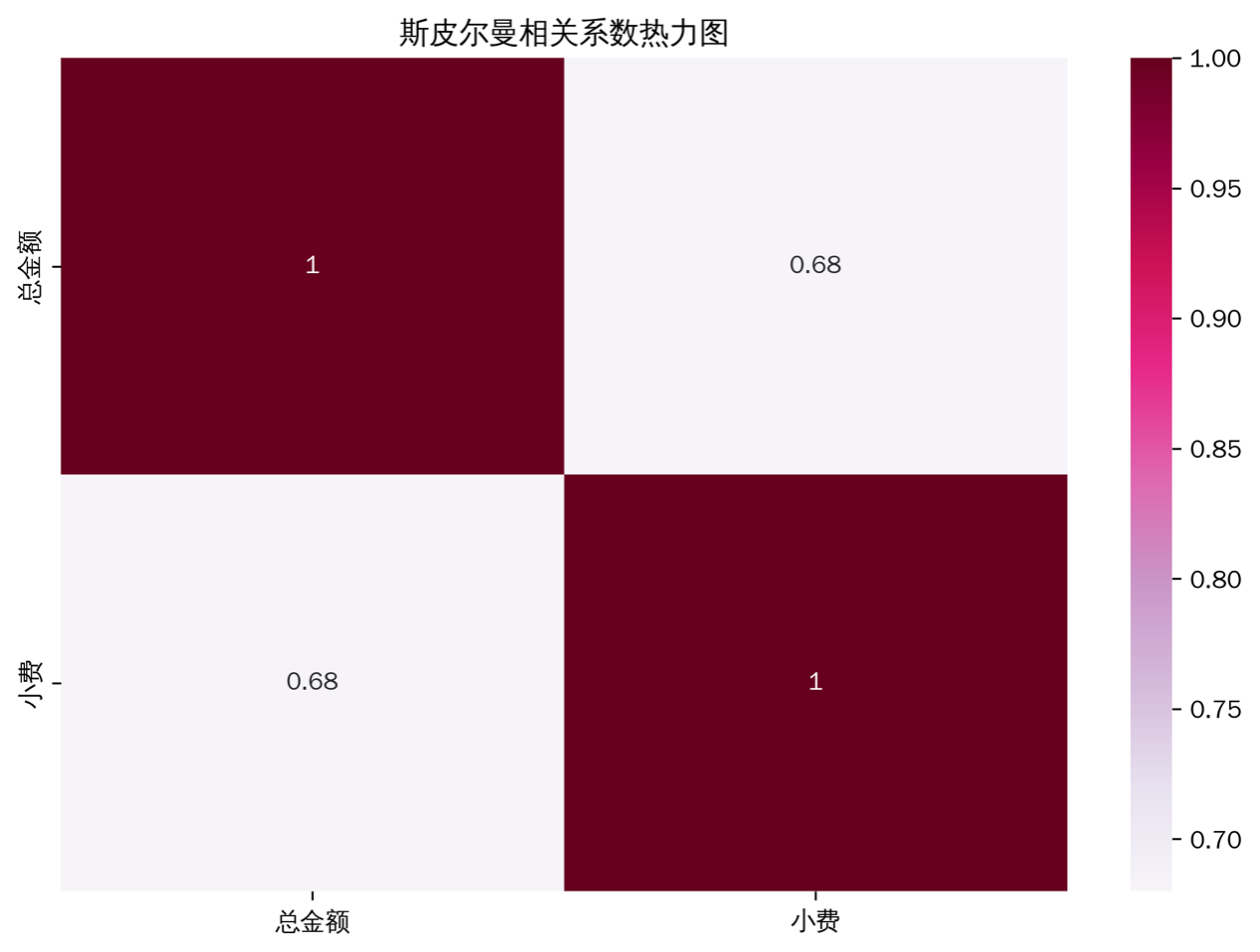
(1)斯皮尔曼相关系数的运用

**斯皮尔曼相关系数(Spearman`s rank correlation coefficient)**是一种用于评估两个变量之间的单调关系的统计量。它主要用于测量两个变量之间的等级(排名)关联程度，而不是它们的线性关系。斯皮尔曼相关系数的有点在于它不要求数据具有正态分布，并且对于异常值的敏感度较低，X，Y为两组独立分布的数据，其中样本个数为N,和分别表示两张随机变量的第i个值,其中i =1,2,..,N

首先对X和Y进行同时降序或者升序排列，得到两个元素x和y，其中和分别为和在各自集合中的排序。设定集合d为集合X与Y集合中相同为元素排序之差，d集合各元素计算公式如下:

斯皮尔曼计算公式如下:

利用python绘制的总金额和小费的斯皮尔曼相关系数矩阵热力度图如下:



由图可以看出总金额和小费之间具有较强的正相关关系

(2)斯皮尔曼相关系数检验

本文对于总金额和小费相关系数检验步骤如下：

***Step1:***提出假设

原假设:斯皮尔曼相关系数r=0;

备择假设:斯皮尔曼相关系数 r ！= 0;

设定置信水平为95%

***Step2***:计算p值

本文利用python进行斯皮尔曼相关系数检验，结果如下：

|  |  |  |
| --- | --- | --- |
|  | P值检验 |  |
|  | 总金额 | 小费 |
| 总金额 | 1.00 | 2.501158440923619e-34 |
| 小费 | 2.501158440923619e-34 | 1.00 |

表3

由表3可知，总金额和小费之间的p值可以近似为0，所以在95%的置信水平上拒绝原假设，可以认为总金额和小费之间存在显著相关性

**5.2问题二：分析男性顾客和女性顾客对于小费的慷慨度**

思路：想办法判断两者之间是否有显著性差异，如果有就有慷慨度之差，没有就看谁的离散程度大，谁的慷慨度就高

**5.2.1独立样本t检验**

对于两个独立样本1和2的t检验，计算t值的公式为:

其中和是样本1和2的均值，和是样本1和2的方差,和 是1和2的样本容量

自由度计算公式为:

本文独立样本t检验步骤如下：  
***Step1***:提出假设

原假设 : 两个样本集之间均值没有显著性差异

备择假设: 两个样本集之间均值有显著性差异

设定置信水平为95%

***Step2***:计算p值

本文利用python计算p值，结果如下所示:

表4

|  |  |  |
| --- | --- | --- |
|  | P值检验 |  |
|  | 男顾客 | 女顾客 |
| 男顾客 | 1.00 | 0.13780683808650296 |
| 女顾客 | 0.13780683808650296 | 1.00 |

从表4可知，t检验的p值约为0.1378大于显著性水平0.05，故接受原假设，男顾客和女顾客没有显著性差异

**5.2.2绘制柱状图**

柱状图绘制如下：

图表, 箱线图

描述已自动生成

图6男顾客和女顾客的小费柱形图

**5.2.3绘制箱型图**

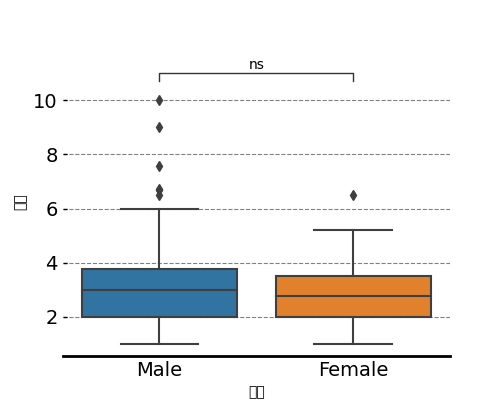


图7男顾客和女顾客小费箱型图

从图中可以看出男顾客的小费中位数高于女顾客，箱体上线边缘分别表示第一四分位数和第三四分位数、男顾客的四分位距比女顾客的四分位距要大，表明男顾客的小费分布比较分散。男顾客的离群值比女顾客多，说明在极端情况下，男顾客的小费金额高于女性顾客。

综上所述，性别对慷慨度没有显著性影响，极端情况下，男顾客更慷慨。

**5.3问题三：分析日期和总金额之间的关系**

利用python代码的索引功能将原始数据分为四类，一类周四，一类周三，一类周六，一类周日，再利用python绘制箱型图，每个样本之间两两进行独立样本t检验，比较两个独立样本的均值是否存在显著差异，进而分析日期和小费的关系

**5.3.1进行独立样本t检验**

原理和问题2相同

***Step1***:提出假设

原假设 : 两个样本集之间均值没有显著性差异

备择假设: 两个样本集之间均值有显著性差异

设定置信水平为95%

***Step2***:计算p值

我们对这四类数据两两排列组合，统计到的p值如下：

表5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | P值检验 |  |  |
|  | 周天 | 周六 | 周五 | 周四 |
| 周天 | 1.00 | 0.500676892195244 | 0.05797254510953542 | 0.009901038244588888 |
| 周六 | 0.500676892195244 | 1.00 | 0.13831793863570302 | 0.05517586159734875 |
| 周五 | 0.05797254510953542 | 0.13831793863570302 | 1.00 | 0.8068054241094087 |
| 周四 | 0.009901038244588888 | 0.05517586159734875 | 0.8068054241094087 | 1.00 |

依据表5的数据，我们只能接受周天和周四，周六和周四，周天和周五在0.05的显著性指标下是可以认为示有显著差异的，拒绝原假设；其他的组合就是接受原假设，没有显著差异性

**5.3.2绘制箱型图**

图表, 箱线图

描述已自动生成

图8总金额和日期关系箱型图

通过图8可得：

1. **中位数**：箱型图的中间线表示数据的中位数，由图八可以看出，周天的总金额中位数高于其他日期；
2. **四分位距**：箱型图的上下边缘分别表示第一四分位数(Q1)和第三四分位数(Q3)。周日总金额的四分位距比其它日期的四分位距要打，表明周日顾客的总金额分布较为分散
3. **离群值**：图中圆形点表示离群值，在周四顾客的小费离群值最多，这说明在极端情况下，顾客在周四消费的总金额最高
4. **显著性检验**：图中展现了这四类数据的显著性结果

周四和周日之间存在显著性差异(\*\*);

周四和周六之间存在显著性差异(\*);

周五和周日之间存在显著性差异(\*)。

基于分析，周四与周日和周六都有显著性差异，且周四的离群值最多，说明在周四的顾客消费总金额最多，餐厅可以在这一天进行优化宣传

**5.4问题四：分析性别+抽烟的组合因素对慷慨度的影响**

**5.4.1 绘制箱型图**

使用python索引加上字符串拼接的方式来实现性别+抽烟的组合因素，对于0\_1 第一个数字0或者1表示的是女性还是男性，第二个数字0或者1表示不抽烟和抽烟

本文使用python绘制箱型图:

图表, 箱线图

描述已自动生成

图9性别+抽烟和小费关系的箱型图

依据图9的信息，

1. **离散值**：对于女性之间的比较，抽烟的女性更加慷慨；对于男性之间的比较，抽烟的男性比不抽烟的更慷慨
2. **显著性分析**：

四个全是ns

显然，四组数据之间都没有显著性(ns)，因此我们可以认为，性别+抽烟的组合因素与小费之间无显著性差异，但是依据离散值情况，我们还是可以认为，在极端情况下，抽烟的人(无论男女)比不抽烟的要慷慨

**5.5问题五：分析用餐时间对小费的影响**

**5.5.1 绘制箱型图:**

**图表, 箱线图

描述已自动生成**

图10用餐时间与小费的箱型图

1. **中位数**：Dinner的中位数比Lunch时间段的高，说明它的小费中位数也比Lunch高
2. **四分位距**：Dinner的上下边栏间高度比Lunch的高，表示Dinner的顾客的小费分布较为分散(有大面额，也有小面额)
3. **离散值**：Dinner的离散值显著比Lunch高，说明在极端情况下，顾客在Dinner餐段的小费消费最高
4. **显著性分析**：

Dinner和Lunch餐段时间内的小费分布有显著的差异(\*)

综上，顾客在Dinner餐段给的小费比Lunch餐段的小费要多，两者之间有着显著性差异。

1. **模型的优化以及推广**

本文使用的主要是斯皮尔曼相关系数法，是由于我们遇见的数据是非正态分布的，如果我们遇见了符合JB正态检验假设下的服从正态分布的数据，我们可以考虑使用**皮尔逊相关系数**，这个系数计算的结果和斯皮尔曼相关系数计算正态分布下数据的速度要更快，结果会更优，同理，二者也可以在同一正态分布数据得到的结果取均值来更加准确的逼近真实结果

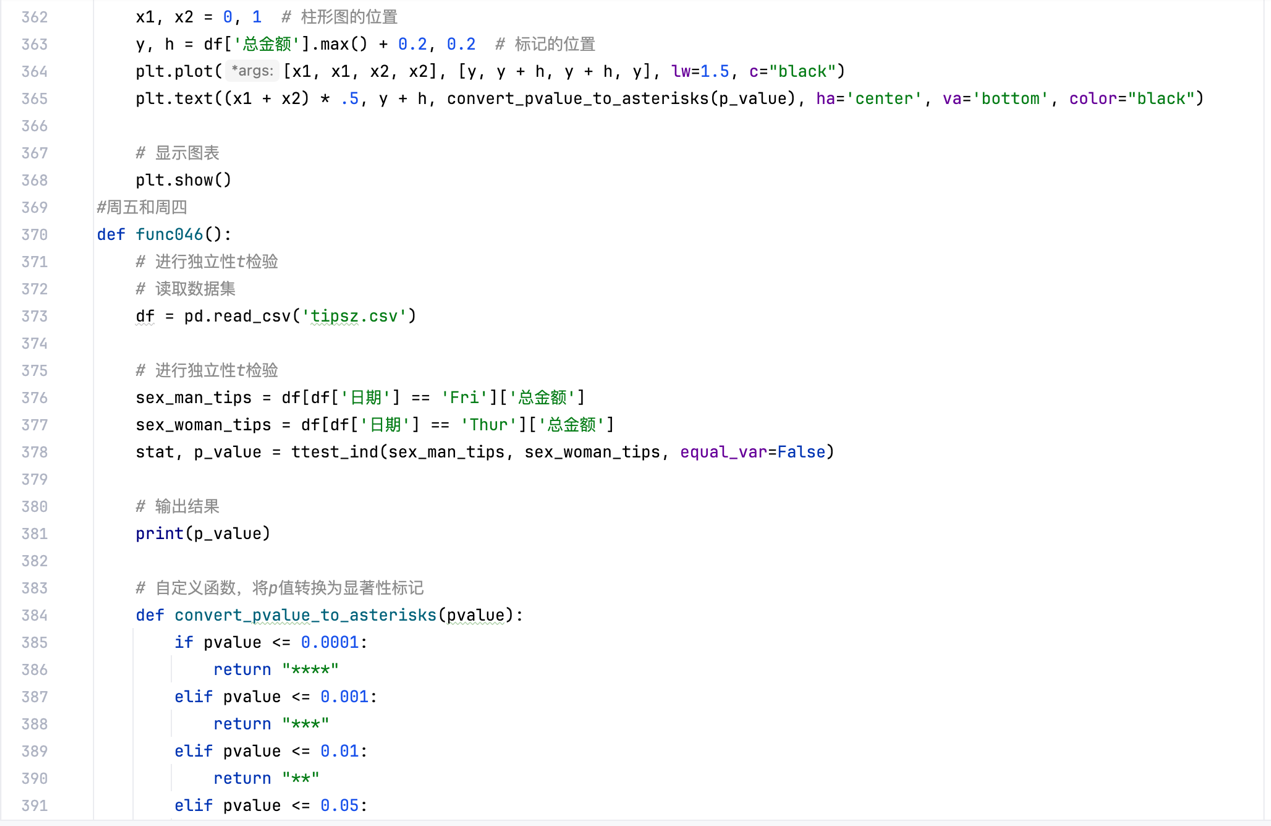
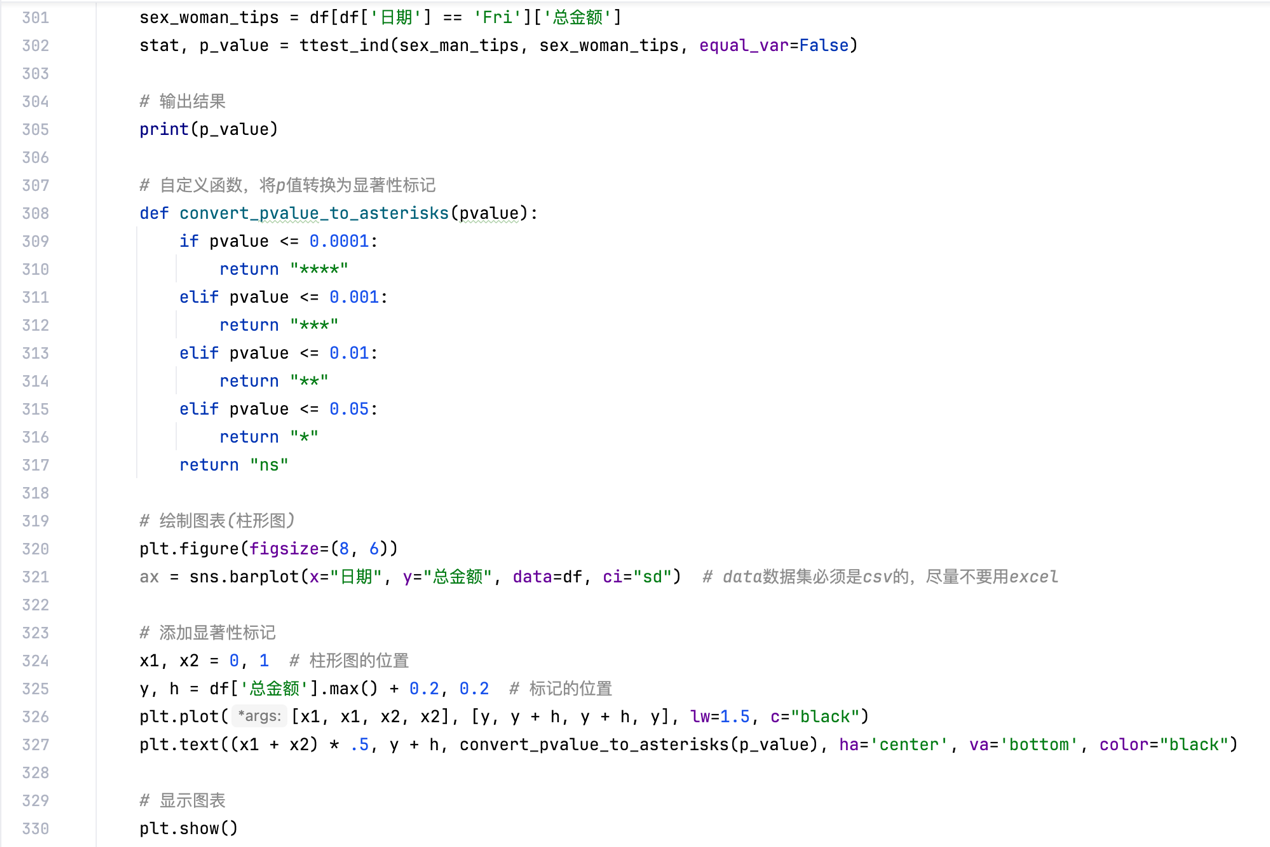
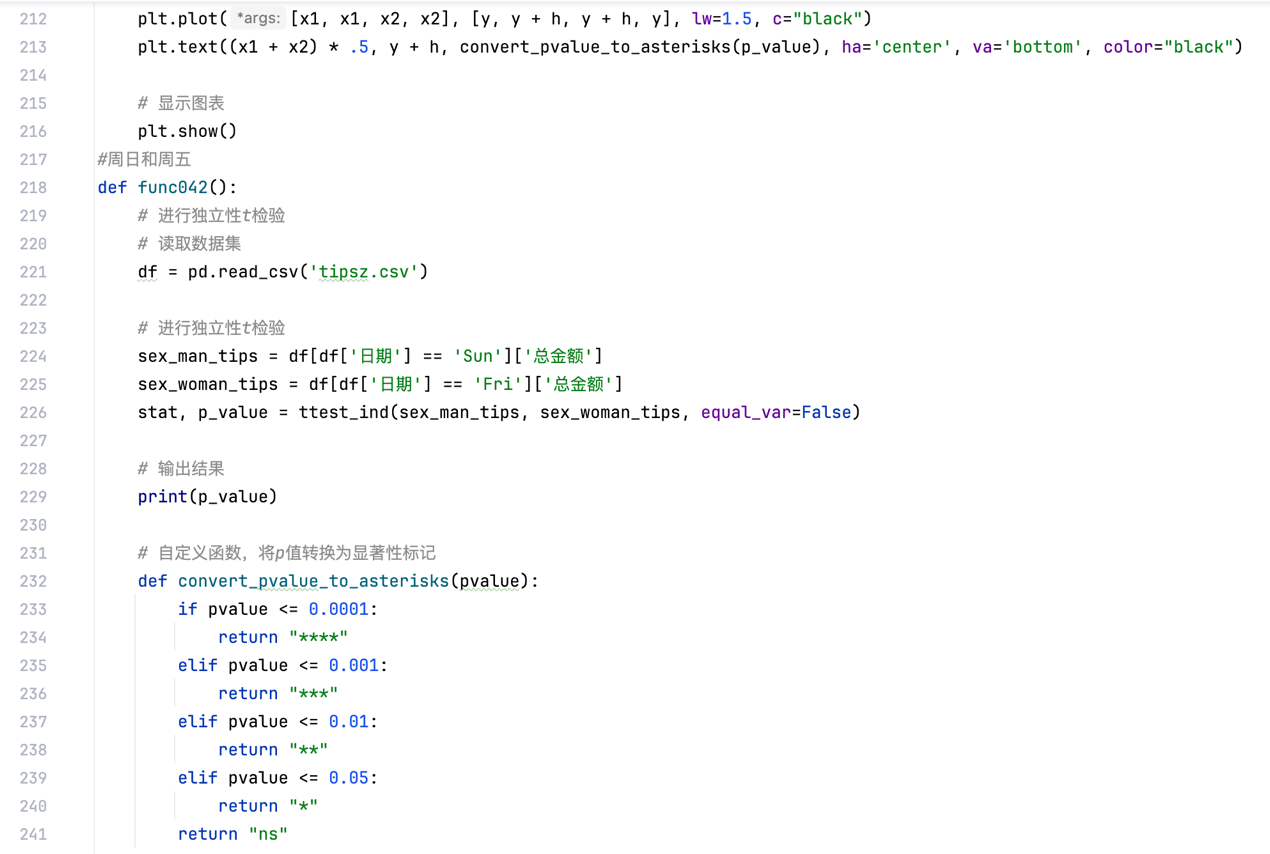
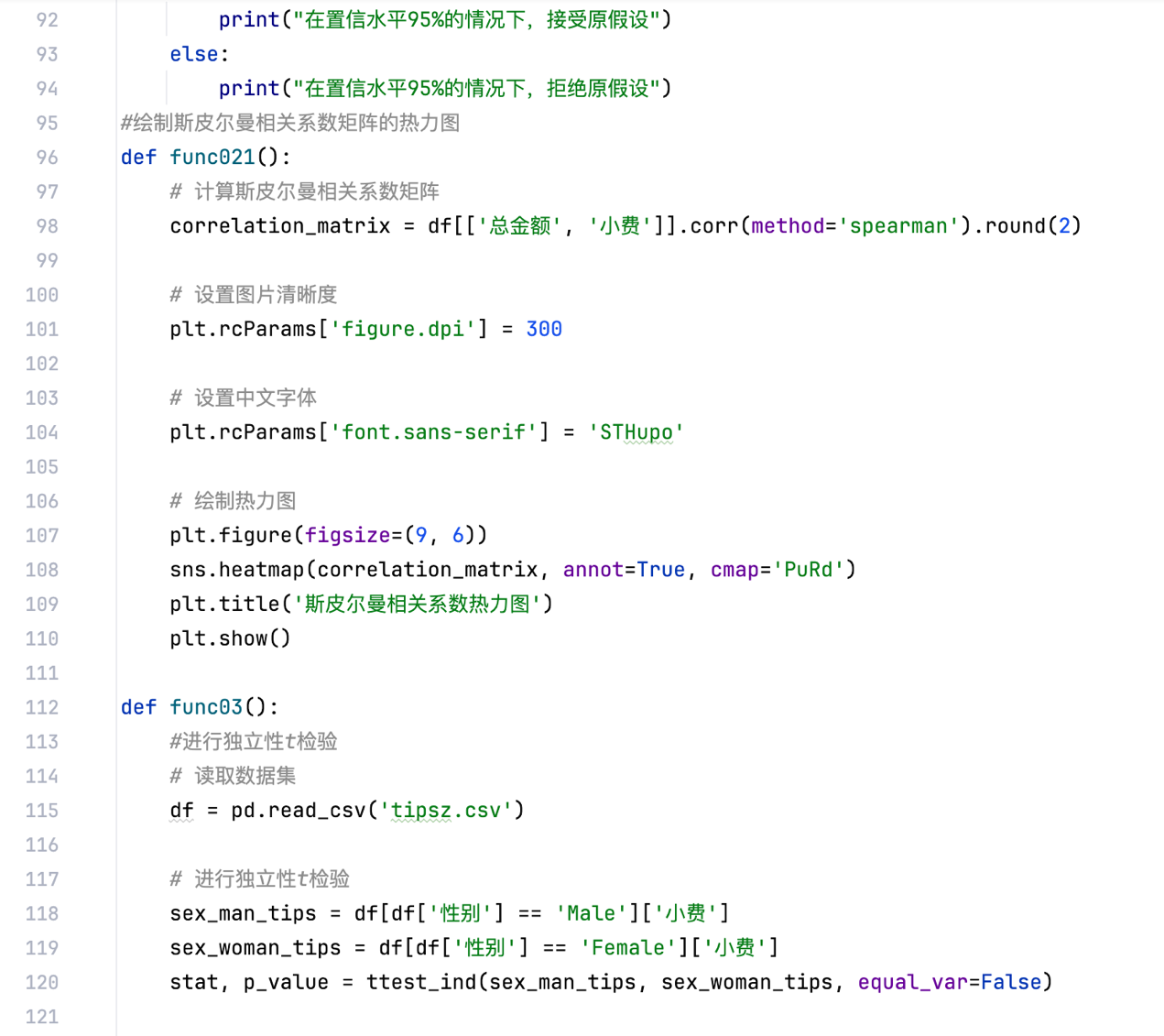
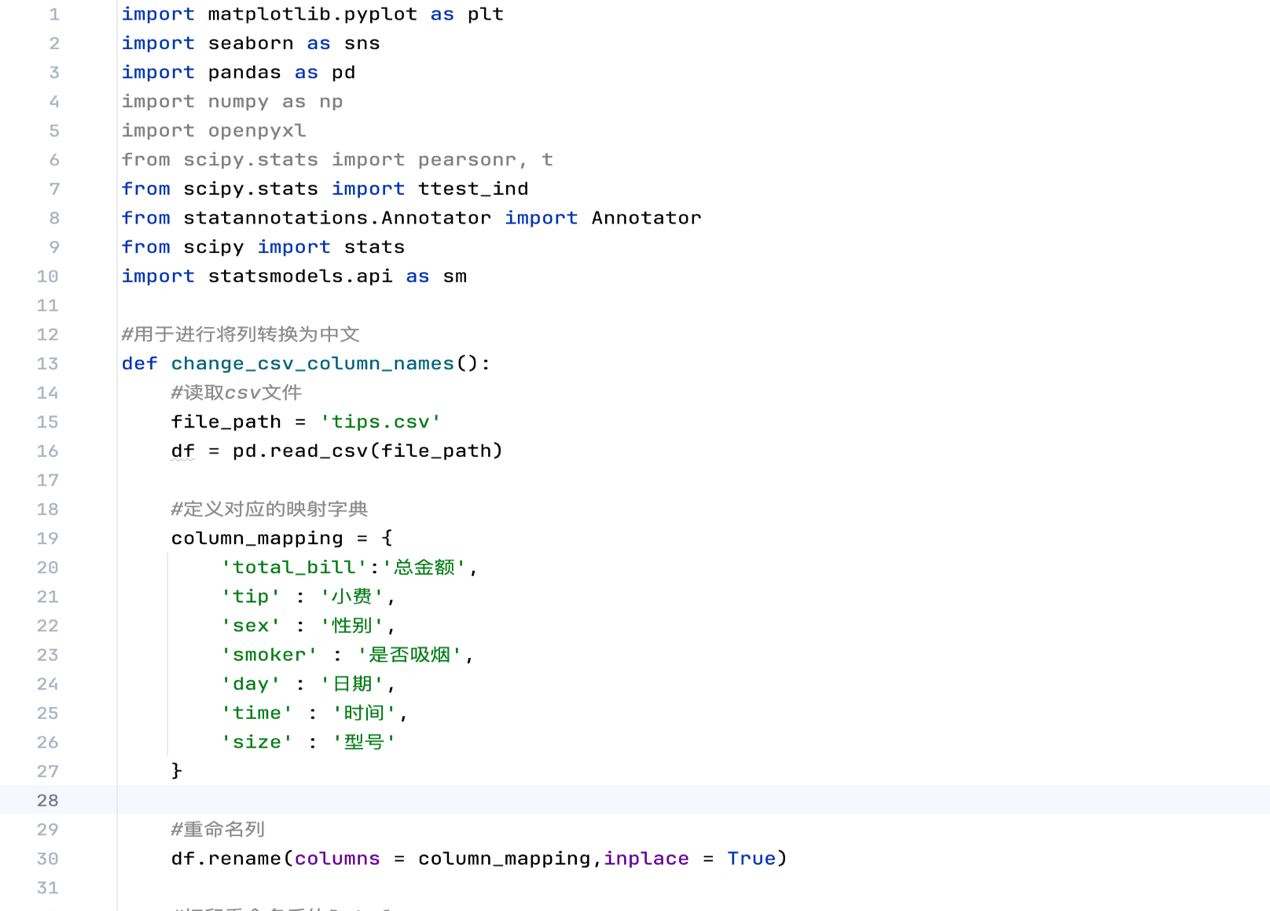
1. **参考文献**

[1]wiki百科，斯皮尔曼相关系数spearman，<https://baike.wiki.com/item/spearman，2024>.

[2]司守奎、孙玺箐，《数学建模算法及其应用》，《python数学实验与建模》，《python高等数学实验》.[M].北京;科学出版社,2020.

[3]wiki百科，独立样本t检验，[https://baike.wiki.com/item/t.test ,2024](https://baike.wiki.com/item/t.test%20,2024).

1. **附录**

****