**项目报告**

**项目：** 基于R语言的空气质量数据分析与可视化研究

**项 目 报 告**

|  |  |
| --- | --- |
| **实验名称** | 基于R语言的空气质量数据分析与可视化研究 |
| **实验目的** | 本实验旨在综合运用本学期所学的数据分析理论知识与R语言编程技能，针对城市空气质量监测数据进行系统性研究。通过对数据的采集、清洗、整理与可视化，探索污染物浓度与气象条件之间的关联特征，深入挖掘数据背后的潜在价值。通过规范、熟练地使用R语言实现数据预处理、统计分析和可视化呈现，并最终整理成一份完整、严谨的数据分析报告，为环境治理和空气质量预测提供科学依据。 |
| **实验原理** | 一、描述性统计分析  描述性统计分析是数据分析的基础，通过计算数据的中心趋势、离散程度、偏态等统计量，帮助我们了解数据的分布特性和整体趋势。  **平均数:** 数据的中心值，表示总体趋势。  **标准差:** 测量数据的离散程度，反映数据的波动性。  **中位数:** 数据排序后居中的值。  **偏态系数:** 衡量数据分布的对称性，偏态系数为0表示对称分布；正偏态表示右偏，负偏态表示左偏。  **四分位数:** 用于衡量数据的分布位置（25%、50%、75%分位数）。  **最大值和最小值:** 数据的极端值。  **缺失值分析:** 检查数据中是否存在缺失值，并分析其分布和可能影响。 **频数分布：**通过将数据分成若干区间，统计每个区间内数据出现的频数，从而形象化地展示数据的分布特征。  **相关系数：**相关系数是衡量两个变量之间线性关系强弱和方向的统计量，其取值范围为 [−1,1]。  **相关系数的检验：**相关系数的检验用于评估两个变量之间的相关性是否显著。检验的基本思想是通过假设检验判断样本相关系数是否显著不同于零，从而推断总体相关性。  偏相关系数：偏相关系数是衡量两个变量在控制住其他变量的影响后，所具有线性关系强弱的统计量。它在多变量分析中非常重要，用于剔除第三方变量（或多余变量）的干扰，揭示变量间的纯粹相关性。  **边际频数：**边际频数是交叉列联表中行或列的总计数，表示各类别数据的总量，而不区分与其他变量的关联。  **边际百分比：**边际百分比是将边际频数转换为样本总量的百分比，表示各类别在样本中的相对占比。  **两分类型变量相关性检验：**在列联表的基础上，利用列联表数据，分析表中两分类型  变量的总体相关性。采用的方法是卡方检验。  二、数据可视化与图形标注  数据特征可视化的基本工具是统计图形。一方面，图形是直观展示变量分布特征以及变量在不同样本组分布特征差异性的重要工具。另一方面，R的图形绘制功能强大，图形种类丰富，在数据可视化方面优势突出。  **茎叶图：**是一种用于展示数据分布的图形化方法，它将数据按数值范围进行分组，以“茎”代表数据的高位部分，以“叶”代表低位部分。通过这种方式，茎叶图能够直观地展示数据的分布形态，同时保留每个数据点的具体数值。  **箱线图：**是一种用于显示数据分布、集中趋势和离群点的可视化工具。它通过展示数据的 最小值、下四分位数（Q1）、中位数（Q2）、上四分位数（Q3） 和 最大值，帮助我们更好地理解数据的集中程度、分散情况以及异常值。通过图中的箱线图可以很清晰的观测不同污染物浓度的各个指标。  **直方图：**直方图通过将数据划分为若干区间，并显示每个区间内数据的频数或频率，直观地呈现数据的分布形态。直方图可以帮助分析数据的集中趋势、离散程度、偏态、对称性等特征。  **核密度图：**是一种用于估计数据分布的平滑图形方法，它通过对数据点进行平滑处理，生成一个连续的概率密度曲线。核密度图不依赖于区间划分，而是通过核函数平滑数据，能够提供数据分布的更细腻的视图。  **小提琴图：**是一种结合了箱线图和核密度图特点的可视化图表，用于展示数据分布的形状、集中趋势、离散程度以及数据的密度。它的形状类似于小提琴，图中不仅包含了箱线图的核心统计量（如中位数、四分位数），还通过对称的密度曲线展示了数据的分布密度，能够更直观地反映出数据的分布模式和变化趋势。  **克利夫兰图：**克利夫兰图使用点而非长条来表示数据的值，每个点的位置对应一个特定的数值，这种方式使得图表更加简洁、清晰，尤其适合显示精确的数值。  **折线图：**是一种用于显示数据变化趋势的图表，通常用于时间序列数据的可视化。它通过将数据点连接成线段，帮助展示数据随时间、类别或其他变量的变化情况。  简单柱形图：是一种用于展示不同类别之间数量差异的图表，通常通过矩形的柱状条来表示每个类别的数值大小。柱形图的高度或长度与对应类别的数值成正比，适用于比较不同类别的数据。  **饼图：**是一种用于显示各个部分在整体中所占比例的图表，通过将圆形分割成若干扇形区域来表示每个类别的占比。每个扇形的角度大小与该类别在整体中的比重成正比，适用于展示各部分相对大小的情况。  **散点图：**是一种用来展示两个变量之间关系的图表。它通过在二维坐标系中绘制数据点，每个点的横坐标和纵坐标分别表示两个变量的取值。  **高密度散点图：**是一种用于展示大量散点数据分布的可视化图表。它通过调整散点的密度或颜色深浅来表示数据点的密集程度，能够更清晰地展示数据中密集区域和稀疏区域的分布情况。高密度散点图特别适合于展示大规模数据集中的关系和模式。  **矩阵散点图：**是一种用于展示多个变量之间关系的可视化工具，特别适用于探索多个变量之间的相互关系。矩阵散点图通过将每一对变量的散点图排列在一个矩阵中，每个单元格代表两个变量之间的散点图，从而帮助分析不同变量之间的相关性、趋势以及潜在的模式。  **Q-Q图：**是一种用于比较分布的图表，尤其用于检验一个数据集是否符合某个理论分布（如正态分布）。Q-Q图通过将数据的分位数与理论分布的分位数进行比较来进行可视化。如果数据符合该理论分布，那么Q-Q图上的点将沿着一条直线分布，否则就会偏离直线。  **图形注释和标题:** 添加图例、坐标轴标题、数据标签等，提高图表的可读性和直观性。  三、数据分析理论  **1. 均值比较检验**  Levene's Test 检验方差同质性  对来自两个不同总体的两个样本分别计算样本均值。计算各观测与本组样本均值差的绝对值，得到两个绝对离差样本利用单因素方差分析方法，依据F统计量的观测值和概率-P值判断两组绝对离差的均值是否存在显著差异，即：判断两组的平均绝对离差是否存在显著差异。若无显著差异表明两总体方差齐性。  独立样本 t 检验  独立样本 t 检验用于比较两个独立样本的均值是否存在显著差异。  检验步骤：  1. 假设设定：  零假设（H₀）: 两个独立样本的均值相等，即没有显著差异。  备择假设（H₁）: 两个独立样本的均值不相等。  2. 计算 t 统计量  3. 自由度的计算  4. 确定显著性水平α：  一般设定为 0.05。如果 t 值的 p 值小于α，则拒绝零假设，说明两个样本均值存在显著差异。  5. 结论：  如果 p 值小于显著性水平，拒绝零假设，认为两组均值有显著差异。  如果 p 值大于显著性水平，接受零假设，认为两组均值没有显著差异。  功效分析  功效分析是统计学中评估假设检验能力的过程，用于确定在给定条件下，检验能够正确发现效应的概率。它有助于确定所需的样本大小，以保证检验具有足够的能力来检测显著的差异。功效分析的关键要素包括显著性水平、效应量、样本量和检验的功效。  Wilcoxon秩和检验  是一种非参数统计方法，用于比较两个独立样本的分布是否存在显著差异。它不要求数据服从正态分布，适用于处理非正态分布数据或顺序数据。  检验步骤：  1. 假设设定：  零假设（H₀）: 两个样本来自相同的分布，即两个样本的中位数相同。  备择假设（H₁）: 两个样本来自不同的分布，即两个样本的中位数不同。  2. 数据排序：  将两个样本的数据合并并按大小排序。  对每个数据分配一个秩次（从最小值开始排序）。  3. 计算秩和：  计算每个样本的秩和，即样本中所有数据的秩次之和。  4. 计算U统计量：  U统计量用来检验两个样本的秩和是否有显著差异。  U=R−n(n+1)​ /2 其中，R 是样本的秩和，n 是样本大小。  5. 计算p值：  根据U统计量和样本的大小，计算p值。  如果p值小于设定的显著性水平（如0.05），则拒绝零假设，认为两个样本有显著差异。  置换检验  是一种非参数统计方法，用于检验两组或多组数据是否具有显著差异。置换检验不依赖于数据的分布假设，尤其适用于数据不满足正态性假设时。它通过反复随机重新排列数据来生成一个基于无效假设的分布，从而评估观察到的统计量是否显著。  置换检验的步骤：  假设设定：  零假设（H₀）：两组数据的分布没有显著差异。  备择假设（H₁）：两组数据的分布存在显著差异。  计算观察到的统计量  进行置换：  随机将两组数据的标签或数据点进行交换（或将总体数据随机分配到两组）。  重新计算每次置换后的统计量。  重复置换：  多次（通常为数千或更多次）随机置换数据，并计算每次置换后的统计量。  计算 p 值：  p 值是观察到的统计量与置换后的统计量分布的比较结果。通常，p 值为观察到的统计量与置换统计量分布中大于等于观察到的统计量的比例。  如果 p 值小于显著性水平（例如0.05），则拒绝零假设，认为两组数据有显著差异。  **2. 方差分析（ANOVA）**  单因素方差检验  单因素方差分析用来研究一个控制变量的不同水平是否对观测变量产生了显著影响。这里，由于仅研究单个因素对观测变量的影响，因此称为单因素方差分析  方差分析的步骤：  1. 假设设定：  零假设（H₀）: 各组的均值相等，即不同组别之间没有显著差异。  H0​:μ1​=μ2​=⋯=μk​  备择假设（H₁）: 至少有一组的均值与其他组的均值不同。  2. 计算总方差（SST）：  总方差是所有数据点与总均值的偏差平方和，用于衡量数据的总变异性。  3. 计算组间方差（SSB）：  组间方差反映了组均值之间的差异，衡量组别之间的变异。  4. 计算组内方差（SSW）：  组内方差反映了组内个体之间的差异，衡量组内的变异  5. 计算均方（MS）：  组间均方：MSB= SSB​/(k-1) 组内均方：MSW= SSW/(N-k)  6. 计算 F 统计量：  F 统计量用于检验组间均值是否有显著差异：F=MSB​/MSW  7. 计算 p 值：  根据 F 统计量和自由度，计算 p 值。  8. 结论：  如果 p 值小于显著性水平（例如 0.05），拒绝零假设，认为不同组之间有显著差异。  如果 p 值大于显著性水平，接受零假设，认为不同组之间没有显著差异。  单因素协方差分析  用于在比较多个组的均值时控制协变量的影响。协方差分析结合了方差分析和回归分析的优点，能够排除协变量对因变量的影响，从而更准确地检验组间的差异。  协方差分析的步骤  设定假设：  零假设（H₀）：控制协变量后，各组的因变量均值相等。  备择假设（H₁）：控制协变量后，至少有一组的因变量均值不同。  检查协变量的作用：  协变量与因变量应存在显著相关性。  协变量与分组因子之间不存在交互作用。  模型分解： 将因变量的总变异分解为以下几部分：  组间变异：因分组因素导致的因变量差异。  协变量变异：因协变量导致的因变量差异。  误差变异：无法解释的随机变异。  调整因变量：  使用回归分析调整因变量的值，剔除协变量的影响。  调整后的因变量值用于组间比较。  计算 F 统计量：  F 统计量用于检验调整协变量后的组间均值是否存在显著差异。  作出决策：  如果 p 值小于显著性水平（如 0.05），拒绝零假设，认为组间均值存在显著差异。  如果 p 值大于显著性水平，则无法拒绝零假设。  多因素方差分析  是用于研究两个或更多因素（自变量）对一个因变量的影响，尤其是在考虑因素之间是否存在交互作用时。多因素方差分析不仅能评估每个因素的主效应，还可以检测因素之间的交互效应。  在进行多因素方差分析时，通常关注以下几个方面：  主效应：每个因素对因变量的独立影响。  交互效应：两个或多个因素之间相互作用的影响，即因素的组合对因变量的影响是否不同于单独因素的影响。  误差：由随机因素或实验设计外的其他因素引起的变异  **3. 线性回归模型**  线性回归模型是一种用于研究自变量与因变量之间关系的统计方法，通过拟合一条直线来描述这种关系。它可以是简单线性回归（单一自变量）或多元线性回归（多个自变量）。通过最小二乘法来估计回归系数，目的是最小化预测值与实际值之间的误差平方和。线性回归模型的核心假设包括变量间存在线性关系、观测值独立性、误差正态性和同方差性。模型评估指标如R²值和均方误差帮助衡量模型的拟合效果和预测准确性。 |
| **实验数据** | **数据来源：<http://archive.ics.uci.edu/datasets>**  **数据形式：**Excel表  **数据内容：**  Date: 日期（年/月/日）。  Time: 时间（时:分:秒，24小时制）。  CO(GT): 一氧化碳浓度的真实测量值（mg/m³）。  PT08.S1(CO): 一氧化碳传感器的输出信号。  NMHC(GT): 非甲烷总烃的真实测量值（µg/m³）。  C6H6(GT): 苯浓度的真实测量值（µg/m³）。  PT08.S2(NMHC): 非甲烷总烃传感器的输出信号。  NOx(GT): 氮氧化物浓度的真实测量值（µg/m³）。  PT08.S3(NOx): 氮氧化物传感器的输出信号。  NO2(GT): 二氧化氮浓度的真实测量值（µg/m³）。  PT08.S4(NO2): 二氧化氮传感器的输出信号。  PT08.S5(O3): 臭氧传感器的输出信号。  T: 气温（°C）。  RH: 相对湿度（%）。  AH: 绝对湿度（g/m³）。 |
| **实验步骤** | **一、数据准备：**将Excel表格中的城市空气质量数据导入到data数据框，并将值为-200（异常值）的记录设为NA。  **二、数据描述性统计：**  1.利用summary函数计算基本的描述统计量。  2.利用sapply函数实现简化  3.利用is.na函数和colSums函数统计异常值  4.利用table函数编制频数分布表  5.利用cor函数计算几个变量的相关系数  6.利用cor.test函数进行变量相关系数的检验  7.利用cor2pcor函数计算偏相关系数  8.利用table函数编制列联表，利用margin函数计算边际频数，利用prop.table函数计算边际百分比，利用chisq.test函数进行卡方检验。  **三、数据可视化**  1.利用stem函数绘制T(温度)的茎叶图  2.利用boxplot函数绘制箱线图  3.利用hist,density函数绘制直方图与核密度图  4.利用vioplot函数绘制小提琴图  5.利用dotchart函数绘制克利夫兰图  6.利用plot函数绘制折线图  7.利用barplot函数绘制简单柱形图  8.利用pie函数绘制饼图  9.利用plot函数绘制散点图  10.利用smoothScatter函数 绘制高密度散点图  11.利用pairs和scatterplotMatrix函数绘制矩阵散点图并添加回归线。  12.利用qqnorm函数绘制Q-Q图  四、数据分析  1. 均值比较检验  (1).利用leveneTest函数进行早上六点和晚上六点CO(GT)的方差同质性检验  (2).利用t.test函数进行两独立样本均值比较检验  (3).利用pwr.t2n.test函数进行功效分析  (4).利用wilcox.test函数进行秩和检验  (5).利用oneway\_test函数进行置换检验  2.方差分析  (1).利用aov,anova函数进行单因素方差分析  (2).利用aov函数进行单因素协方差分析  (3).利用aov函数进行多因素方差分析  3. 线性回归模型  (1).利用lm函数构建线型回归模型，coefficients函数提取回归系数  (2).利用summary函数显示线性回归分析结果的摘要；confint函数显示回归系数默认95%的置信区间  (3).利用predict函数做预测  (4).利用hatvalues函数计算高杠杆值  (5).利用rstudent函数探测离群点  (6).利用cooks.distance函数探测强影响点  (7).利用vif函数计算方差膨胀因子 |
| **分析与讨论** | **数据描述性统计：**  **1.** **利用summary函数计算基本描述统计量**  表1 不同污染物的描述统计量    Min: 最小值1st Qu.: 第一四分位数 Median: 中位数 Mean: 均值3rd Qu.: 第三四分位数 (75%的数据点小于该值)Max: 最大值NA's：缺失值  **2. 利用sapply函数实现简化**  表2 不同污染物的描述统计量    Mean# 平均值SD # 标准差Median # 中位数Skewness # 偏态系数Min # 最小值Max# 最大值Quantiles# 四分位数NA\_Count# 缺失值计数  **3.** **利用is.na函数和colSums函数统计异常值**  表3 不同污染物指标的缺失值数量    **4.** **计算分类变量的频数分布**  表4 CO(GT)污染物的频数分布    **5.** **选择五个变量并计算它们的相关系数矩阵**  表5 相关系数矩阵    O(GT) 与其他变量的相关性：  CO(GT) 与 NMHC(GT): 相关系数为 0.8872，说明 CO 和 NMHC (非甲烷总烃) 之间有强正相关关系。它们的浓度通常是一起变化的，这可能意味着它们在大气中源自相似的污染源，或者它们的生成机制有相似之处。  CO(GT) 与 C6H6(GT): 相关系数为 0.9727，这是一个非常强的正相关，表明 CO 和苯 (C6H6) 的浓度变化非常一致。苯通常与燃料燃烧或车排放相关，因此 CO 和苯的相关性强也符合这一点。  CO(GT) 与 NOx(GT): 相关系数为 0.9513，说明 CO 和 NOx (氮氧化物) 浓度之间有很强的正相关。NOx 是汽车和工业排放的主要污染物之一，与 CO 在很多情况下有共同的来源，因此它们的浓度可能同步变化。  CO(GT) 与 NO2(GT): 相关系数为 0.8614，表示 CO 和 NO2 (二氧化氮) 之间有较强的正相关关系。CO 和 NO2 都是由燃烧过程释放的污染物，因此它们通常在同一时间段内呈现出较强的相关性。  NMHC(GT) 与其他变量的相关性：  NMHC(GT) 与 C6H6(GT): 相关系数为 0.8980，表明 NMHC 和 C6H6 之间有很强的正相关关系。非甲烷总烃 (NMHC) 包含了苯、甲烷等其他挥发性有机物，因此它们之间的高相关性是可以理解的。  NMHC(GT) 与 NOx(GT): 相关系数为 0.8112，表明 NMHC 和 NOx 之间有较强的正相关性。虽然 NMHC 和 NOx 来自不同的污染源，但两者都受到交通排放、工业活动等因素的影响，因此它们之间可能存在一定的关联。  NMHC(GT) 与 NO2(GT): 相关系数为 0.7281，这个相关性较高，但相对较弱。尽管 NMHC 和 NO2 都受到交通排放的影响，它们之间的关系不如 CO 或 C6H6 之间的关系那么紧密。  C6H6(GT) 与其他变量的相关性：  C6H6(GT) 与 NOx(GT): 相关系数为 0.9273，表示 C6H6 和 NOx 之间有非常强的正相关。这可能是因为二者都来自于汽车排放和燃料燃烧的过程中，因此它们的浓度变化表现出较强的一致性。  C6H6(GT) 与 NO2(GT): 相关系数为 0.8467，这个值表示 C6H6 和 NO2 之间也有较强的正相关性。苯和二氧化氮往往同时出现在污染源排放的结果中，因此它们的浓度变化通常呈现正相关。  NOx(GT) 与 NO2(GT): 相关系数为 0.8574，这也是一个非常强的正相关。NOx 和 NO2 都是由燃烧源（如交通、工业）产生的污染物，因此它们的浓度变化密切相关。NOx 主要由氮氧化物组成，其中一部分会转化为 NO2。  **6. CO(GT)和NMHC(GT)变量相关系数的检验**  表6变量相关系数的检验    **t** = 58.083, df = 888  这是 t 统计量的值和自由度（df）。t 统计量用来检验相关系数是否显著不同于零，自由度通常是样本大小减去 2（即n−2）。  p-value < 2.2e-16  p 值非常小，远小于常见的显著性水平（如 0.05 或 0.01）。这表明我们可以拒绝零假设，即认为变量 var1 和 var2 之间的相关性显著不为零。  alternative hypothesis: true correlation is not equal to 0  这是对零假设的对立假设，即检验的目的是判断变量之间的真实相关性是否为零。  95 percent confidence interval: 0.8751908 0.9026709  这个置信区间表示在 95% 的置信水平下，变量 var1 和 var2 之间的真实相关系数落在 0.8751908 和 0.9026709 之间。由于这个区间完全在正相关范围（0 到 1）内，可以认为这两个变量之间有显著的正相关性。  sample estimate: cor = 0.8897342  这是样本数据中计算得到的 Pearson 相关系数，值为 0.8897，说明变量 var1 和 var2 之间有很强的正相关性  **7. CO(GT), NMHC(GT), NO2(GT)偏相关系数**  表7偏相关系数    var1 和 var2 的相关系数是 0.7491这表明 var1 和 var2 之间存在 较强的正相关关系。它们的值变化趋势相似，说明这两个变量可能在某些方面受到相同因素的影响。  var1 和 var3 的相关系数是 0.6828这表示 var1 和 var3 之间存在 较强的正相关性，虽然相关性不如 var1 和 var2 强，但它们之间依然呈现出较为明显的正向关系。  var2 和 var3 的相关系数是 -0.1599这个值接近于零，说明 var2 和 var3 之间几乎没有线性相关性。事实上，负值虽然存在，但相关性非常弱，表明它们之间的关系几乎可以忽略。  **8. 编制列联表 计算边际频数 边际百分比 进行卡方检验**  表8边际频数    在 RH\_binned 的 3 个区间中（Low、Medium、High）相对湿度的总频数：  Low: 2222 次Medium: 4484 次High: 2285 次  相对湿度为 "Medium"（中等湿度）的样本最多，其次是 "Low"（低湿度），最后是 "High"（高湿度）。  在 AH\_binned 的 3 个区间中（Low、Medium、High）绝对湿度的总频数：  Low: 3262 次Medium: 4648 次High: 1081 次  绝对湿度为 "Medium"（中等湿度）的样本最多，其次是 "Low"（低湿度），最后是 "High"（高湿度）。  表9 边际百分比    相对湿度低（Low）和中（Medium）区间的样本大部分对应中等的绝对湿度（Medium）。  当相对湿度为高（High）时，绝对湿度倾向于更高的区间，56.41% 的样本位于中等绝对湿度区间。  对于低绝对湿度（Low），相对湿度偏向低和中等区间，低相对湿度（Low）占比最高。  对于中等绝对湿度（Medium），相对湿度则较为均衡，低和中等相对湿度的比例相当。  对于高绝对湿度（High），相对湿度倾向于中等和高区间  表 10 卡方分布  卡方统计量（X-squared）为 131.6，表示观察数据与期望数据之间存在较大的差异。  自由度（df）为 4，符合 3x3 的列联表维度。  p值非常小（< 2.2e-16），显著性极高，因此我们可以拒绝零假设，认为相对湿度（RH）和绝对湿度（AH）之间具有显著的统计相关性。  因此，结论是 RH 和 AH 之间确实存在显著的依赖关系，说明这两个变量之间的变化不是独立的，湿度水平的变化在一定程度上是相关的。  **数据可视化：**  **1.** **绘制T(温度)的茎叶图**    图1 T(温度)的茎叶图  这段数据展示了温度（T）的茎叶图，通过将每个数据点的整数部分（茎）和小数部分（叶）分开，展示了温度值的分布。数据从负数到较高温度值不等，反映了温度在不同区间内的密集程度。  **2. 箱线图**    图2不同污染物的箱线图  箱线图（Boxplot）是一种用于显示数据分布、集中趋势和离群点的可视化工具。它通过展示数据的 最小值、下四分位数（Q1）、中位数（Q2）、上四分位数（Q3） 和 最大值，帮助我们更好地理解数据的集中程度、分散情况以及异常值。通过图中的箱线图可以很清晰的观测不同污染物浓度的各个指标。  **3. 直方图与核密度图**    图3 T,RH,AH的直方图与核密度图  直方图通过将数据分成多个区间并绘制每个区间的频数（或频率）来显示数据的分布。它可以清楚地展示数据的集中趋势、离散程度以及是否存在异常值。  核密度图是一种平滑化的直方图，它通过计算每个数据点的概率密度，并使用核函数对数据进行平滑，生成一个连续的密度曲线。核密度图展示了数据的 估计概率密度函数，提供了对数据分布形态的平滑视图。  **4. 小提琴图**    图4 T,RH,AH的小提琴图  小提琴图结合了箱线图和密度图的优点，能够有效展示数据的分布情况、集中趋势以及变异性。  **5.克利夫兰图**    图5 T,RH,AH的克利夫兰图  **6. 折线图**    图6 T,RH,AH的折线图  **7. 简单柱形图**    图7 CO(GT)的折线图  **8. 饼图**    图8 CO(GT)的饼图  通过饼图可以直观地看到一氧化碳浓度大部分集中于1-3 mg/m³  **9.散点图**    图9 RH的散点图  **10.** **高密度散点图**    图10 RH的高密度散点图  当观测样本量较大时，所绘制的散点图可能会出现数据点非常集中，有很多数据点重叠在一起的现象。这样的散点图称为高密度散点图。  **11. 矩阵散点图**    图11 T, RH,AH的矩阵散点图  **12 .T, RH,AH的Q-Q图**    图12 T, RH,AH的Q-Q图  **数据分析：**  **1.** **6:00 AM 和 6:00 PM 时段的 CO(GT)数据均值比较检验**  表 11 方差同质性检验    表 12独立样本 t 检验    表 13功效分析    表14 Wilcoxon秩和检验    表15 置换检验    1. Levene's Test 检验方差同质性  我们进行了 Levene's Test 检验方差同质性，结果显示 p 值小于 0.05 (p-value < 2.2e-16)，这意味着我们拒绝零假设，认为两个时间段的 CO(GT) 浓度的方差存在显著差异。因此，后续的 t 检验应该采用不等方差的检验。  2. 独立样本 t 检验  我们进行了 Welch's t-test，其假设为：  零假设：6:00 AM 和 6:00 PM 的 CO(GT) 浓度均值相等。  备择假设：6:00 AM 和 6:00 PM 的 CO(GT) 浓度均值不相等。  t 检验的结果为：t 值：-23.561自由度：368.61 p 值：p-value < 2.2e-16  由于 p 值非常小，远小于 0.05，我们拒绝零假设，认为 6:00 AM 和 6:00 PM 的 CO(GT) 浓度均值存在显著差异。置信区间为 (-2.72, -2.30)，说明 6:00 PM 的浓度显著高于 6:00 AM。  3. 效应量 (Cohen's d)  为了进一步评估均值差异的实际意义，我们计算了 Cohen's d，效应量为 1.81。根据 Cohen's d 的标准，效应量 1.81 表示这是一个大效应，即 6:00 AM 和 6:00 PM 之间的均值差异在实际意义上非常显著。  4. 功效分析  最后，我们进行了功效分析以评估检验的敏感性。根据计算结果：  样本量：6:00 AM 组为 320，6:00 PM 组为 333。  Cohen's d：1.81。  显著性水平：0.05。  功效值：power = 1。  功效为 1，表示该检验具有 100% 的概率在样本量下发现真实的效应。因此，样本量和效应量均足够强，能够充分检测到两组数据之间的差异。  5. Wilcoxon秩和检验  在进行 Wilcoxon 秩和检验 时，提出的假设如下：  零假设（H₀）：6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度分布相同，没有显著差异。  备择假设（H₁）：6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度分布不同，存在显著差异。  W 值：检验统计量为 5480.5，表示在两个时间点（6:00 AM 和 6:00 PM）之间的秩差总和。  p 值：p-value < 2.2e-16，这个值非常小，远小于常见的显著性水平 0.05。  根据得到的 p 值，我们拒绝了零假设，认为 6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度分布存在显著差异。这表明两个时段的 CO(GT) 浓度在统计上不相等，且这种差异具有显著性。  6.置换检验  零假设（H₀）：6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度分布相同，即两组数据的均值无显著差异。  备择假设（H₁）：6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度分布不同，即两组数据的均值存在显著差异。  Z 值：检验统计量为 -17.155，表示两组数据之间差异的标准化量度。  p 值：p-value < 1e-04，这个非常小的 p 值意味着我们有强烈的证据拒绝零假设。  根据 p 值 < 1e-04，我们可以拒绝零假设，认为 6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度分布存在显著差异。这表明，6:00 PM 时段的 CO(GT) 浓度与 6:00 AM 时段相比存在统计上显著的不同。  **2.方差分析**  表16 单因素方差分析    Df (自由度)：  AH：自由度为 1，表示我们对一个因素进行比较。  Residuals（残差）：自由度为 7342，表示样本中其他因素的变异性。  Sum Sq (平方和)：  AH：平方和为 35.7，表示因子（时间段）导致的总变异。  Residuals：平方和为 15116.2，表示样本内部（误差）的变异。  Mean Sq (均方)：  AH：均方为 35.722，表示因子导致的变异与其自由度的比值。  Residuals：均方为 2.059，表示误差与自由度的比值。  F value (F 值)：计算出的 F 值为 17.35，它衡量了因子组之间的变异相对于组内误差的变异。如果 F 值较大，通常意味着组之间存在显著差异。  Pr(>F) (p 值)：p 值 = 3.144e-05，非常小的 p 值（小于 0.05）表明我们可以拒绝零假设，即 6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度存在显著差异。  零假设与备择假设  零假设（H₀）：6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度均值相等。  备择假设（H₁）：6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度均值不相等。  结论  根据方差分析的结果：  F 值 = 17.35 和 p 值 = 3.144e-05 表明我们有足够的证据拒绝零假设。  因此，我们可以得出结论：6:00 AM 和 6:00 PM 时段的 CO(GT) 浓度之间存在显著差异。  表17 单因素协方差分析    AH\_clean（协变量）：  F 值 = 18.36，p 值 = 1.85e-05：协变量 AH\_clean 对 CO(GT) 浓度有显著影响。小的 p 值（小于 0.05）表明我们可以拒绝零假设，认为 AH\_clean 在 CO(GT) 浓度的变化中起到了显著作用。  group（自变量）（Time 的值在 18:00:00 到 22:00:00 之间）：  F 值 = 429.13，p 值 < 2e-16：分组变量 group（即实验组与对照组的差异）对 CO(GT) 浓度也有显著影响。p 值极小，表明 group 的作用非常显著。  Residuals（误差项）：  误差项的平方和为 14281，均方为 1.9，F 值为 1.9，说明剩余变异较小。  结论  AH\_clean 对 CO(GT) 浓度的影响是显著的，说明协变量 AH\_clean 与 CO(GT) 浓度之间存在明显的关系。可以推测 AH\_clean 可能与 CO(GT) 浓度之间有直接或间接的联系。  group 变量的影响也非常显著，F 值高达 429.13，说明实验组和对照组之间的差异是显著的。这表明实验组与对照组在 CO(GT) 浓度上有显著差异。  从 p 值 小于 0.05 的结果来看，协方差分析表明 AH\_clean 和 group 这两个因素都对 CO(GT) 浓度产生了显著影响。  表18 多因素方差分析    模型：  CO(GT)∼T+RH+AH  我们检验温度（T）、相对湿度（RH）和空气湿度（AH）对 CO(GT) 的影响。  1. 温度（T）：  F 值 = 100.035，p 值 < 2.2e-16：温度对 CO(GT) 浓度有显著影响。极小的 p 值（远小于 0.05）表明我们可以拒绝零假设，温度对 CO 浓度有显著影响。  2. 相对湿度（RH）：  F 值 = 47.558，p 值 = 1.064e-11：相对湿度也对 CO(GT) 浓度有显著影响。p 值远小于 0.05，说明 RH 是影响 CO(GT) 浓度的一个重要因素。  3. 空气湿度（AH）：  F 值 = 15.735，p 值 = 7.921e-05：空气湿度对 CO(GT) 浓度也有显著影响。p 值远小于 0.05，表示 AH 对 CO 浓度有显著影响，尽管它的影响力小于 T 和 RH。  4. 误差项（Residuals）：  Sum Sq = 1369.26，Mean Sq = 1.664：误差项的平方和（Sum of Squares）为 1369.26，均方为 1.664。这表示数据中未能解释的变异。  结论：  温度（T） 和 相对湿度（RH） 对 CO(GT) 浓度的影响非常显著，p 值都小于 0.001，表明这两个因素是主要的影响因素。  空气湿度（AH） 也对 CO(GT) 浓度有显著影响，但相对于温度和湿度来说，其影响较小。  由于所有 p 值都非常小（均小于 0.05），我们可以得出结论，这些因素都显著地影响了 CO(GT) 浓度。  **线性回归模型：**  表19回归分析结果对象名    表20 线性回归模型    表21回归系数的置信区间    表22 NO2(GT)的预测值    1. 模型系数分析  模型公式为：NO2(GT)=122.1561−0.3917⋅T−31.2173⋅AH−0.5686⋅Month+2.6616⋅Hour  截距 (Intercept):122.1561  这是当所有自变量（温度 `T`、绝对湿度 `AH`、月份 `Month` 和小时 `Hour`）均为0时，NO2(GT) 的预期值。尽管这个数值可能在实际情况中没有实际意义，但它是回归方程的基线值。  温度 (T):-0.3917  温度的系数为负，表明温度每增加1度，NO2(GT)浓度预期下降0.3917单位。这个影响是显著的，温度对NO2(GT)的影响较大。  绝对湿度 (AH): -31.2173  绝对湿度的系数是负的，表示当绝对湿度每增加1单位，NO2(GT)浓度将下降31.2173单位。这个影响也显著。相对于温度，绝对湿度对NO2(GT)的影响更强。  月份 (Month): -0.5686  月份的系数为负，表示随着月份的增加，NO2(GT)浓度有微小的下降趋势。这个影响显著，但相对较小（p值为0.000223）。  小时 (Hour): 2.6616  小时的系数为正，表示随着小时的增加（时间推移），NO2(GT)浓度会增加。这个影响非常显著。  2. 模型系数的置信区间  根据95%的置信区间，我们可以得到每个系数的可能取值范围：  温度 (T): [-0.5362, -0.2472]  绝对湿度 (AH): ([-34.5603, -27.8744]  月份 (Month): [-0.8704, -0.2667]  小时 (Hour):[2.519, 2.804]  3. 预测结果  对于给定的新数据，我们得到以下预测结果：  T AH 时间 预测 NO2(GT)  20.5 0.80 14:30:00 120.1604  22.3 0.75 15:00:00 123.6778  18.2 0.78 16:30:00 127.0090  4. 结果分析  温度 (T) 对 NO2(GT) 的影响是负向的，即温度升高时，NO2(GT)浓度下降。  绝对湿度 (AH) 对 NO2(GT) 的影响也是负向的，绝对湿度升高时，NO2(GT)浓度下降。  月份 (Month) 对 NO2(GT) 有微弱的负向影响，随着月份的增加，NO2(GT)浓度略有下降。  小时 (Hour) 对 NO2(GT) 的影响为正向，随着时间的推移，NO2(GT)浓度上升。  整体而言，温度和湿度对 NO2(GT) 浓度有显著影响，而时间（小时）和月份的影响相对较小。根据模型和预测结果，可以得出结论，温度和湿度是影响 NO2(GT) 浓度的重要因素。  hatvalues(model) 用于计算回归模型中的 帽子值。帽子值是衡量每个观测值在回归模型中对拟合结果的影响力的指标。  rstudent(model) 用于计算回归模型中的 学生化残差。学生化残差是一种标准化的残差，通过考虑每个数据点对模型拟合的影响来进行调整，因此可以更好地评估每个数据点在回归模型中的异常程度。  cooks.distance(model) 是一种衡量数据点在回归模型拟合过程中影响力的统计量。它综合了该点的 残差 和 杠杆值，用于检测哪些数据点对拟合结果有过度影响。大库克距离表示某个数据点对回归模型的拟合有较大的影响，可能是一个异常值或杠杆点。  表23 线性回归模型的VIF（方差膨胀因子）    1. VIF (方差膨胀因子)  VIF (方差膨胀因子) 是衡量回归模型中每个自变量（预测变量）与其他自变量的多重共线性的指标。VIF 值越大，表示该变量与其他变量的相关性越强，可能存在多重共线性问题。  VIF > 10: 存在严重的多重共线性问题。  VIF > 5: 存在一定的多重共线性问题。  VIF ≈ 1: 自变量之间没有多重共线性问题。  结果分析：  T 的 VIF 值为 1.92，表明温度与其他变量（如湿度、月份和小时）之间的多重共线性比较弱。  AH（绝对湿度）的 VIF 为 2.06，稍高一些，但仍低于 5，说明它与其他变量之间有一定的相关性，但不严重。  Month 的 VIF 为 1.23，几乎没有多重共线性。  Hour 的 VIF 为 1.08，也没有明显的多重共线性。  总体来说，VIF 值都比较低，说明模型中的自变量之间并不存在严重的多重共线性问题。 |
| **思考与问答** | **实验过程中使用的重要知识点**  数据准备  导入数据后，识别并处理异常值（如替换为 NA）是数据清洗的关键步骤。  数据描述性统计  包括计算基本统计量、检测异常值、生成频数分布表以及进行相关分析和卡方检验，帮助快速了解数据特性和变量关系。  数据可视化  通过多种图形（如箱线图、直方图、小提琴图、散点图、矩阵散点图等），直观展示数据分布和变量关系。特殊情况下，可以结合茎叶图和高密度散点图探索数据特性。  数据分析  均值比较：结合方差同质性检验（Levene’s Test）、t 检验、秩和检验等方法，分析不同组间的均值差异。  方差分析：利用单因素或多因素方差分析评估变量的显著性影响，还可引入协变量进行协方差分析。  线性回归：通过线性模型评估变量间的因果关系，提取回归系数，分析残差、探测离群点和影响点，并评估模型性能。  **心得体会**  R语言的学习让我体会到理论知识与实践操作的紧密结合在学习中的重要性。从数据导入到清洗，从统计分析到可视化，再到模型构建，每一步都需要扎实的理论基础。通过将 R 语言的各种函数应用于具体的数据集，强化了对抽象概念的理解。例如，偏相关系数的计算原理较为复杂，但通过实际操作 cor2pcor()，让我更清楚偏相关系数在控制其他变量的情况下如何衡量两变量的独立关系。  实验内容覆盖了数据分析的完整流程，包括数据准备、描述性统计、可视化和模型分析。这种系统化的学习让我逐步养成了完整的思维方式：  实验过程中，大量使用 R 语言的内置函数和包，使得我的编程能力显著提高。  通过实验，我更加体会到统计分析和数据挖掘的深度与广度。描述性统计帮助我们从整体把握数据特征；方差分析揭示了不同组间是否存在显著差异；回归分析不仅揭示变量间的关系，还可以预测未来趋势。  **建议**  理论知识的学习往往较为抽象，可以结合实际案例进行授课。例如，结合城市空气质量或医疗数据案例，从数据导入到模型构建进行全流程讲解，能帮助学生更直观地理解知识点。  建议课堂中更多加入互动环节，比如分组讨论、现场数据分析任务等。不仅能提升学习的趣味性，也有助于学生在协作中激发更多创新思维。  鼓励学生选取实际数据集作为课外项目内容，并结合课堂知识进行分析，能更好地巩固所学技能。 |

**实 验 报 告 附 件**

|  |  |
| --- | --- |
| **实验原始记录：步骤与现象** | #数据准备  # 声明  library(readxl)  # 设置文件路径  file\_path <- "D:\\Rproject\\AirQualityUCI.xlsx"  # 导入 Excel 文件  data <- read\_excel(file\_path)  # 获取除了 'Date' 和 'Time' 列以外的列名  cols\_to\_modify <- setdiff(names(data), c("Date", "Time"))  # 替换这些列中的 -200 为 NA  data[cols\_to\_modify][data[cols\_to\_modify] == -200] <- NA  # 定义要分析的变量  variables <- c("CO(GT)", "NMHC(GT)", "C6H6(GT)", "NOx(GT)", "NO2(GT)", "T", "RH", "AH")  #summary函数  summary(data[variables])  # sapply()函数  # 加载必要的包  if (!require(dplyr)) install.packages("dplyr")  if (!require(moments)) install.packages("moments")  #声明  library(dplyr)  library(moments)  # 使用 sapply() 对每个变量计算描述性统计量  stats <- sapply(data[variables], function(x) {  c(  Mean = mean(x, na.rm = TRUE), # 平均值  SD = sd(x, na.rm = TRUE), # 标准差  Median = median(x, na.rm = TRUE), # 中位数  Skewness = skewness(x, na.rm = TRUE), # 偏态系数  Min = min(x, na.rm = TRUE), # 最小值  Max = max(x, na.rm = TRUE), # 最大值  Quantiles = quantile(x, na.rm = TRUE), # 四分位数  NA\_Count = sum(is.na(x)) # 缺失值计数  )  })  # 打印结果  print(stats)  #统计异常值  colSums(is.na(data))  # 计算分类变量的频数分布  table(data$"CO(GT)")  table(data$"NMHC(GT)")  table(data$"C6H6(GT)")  table(data$"NOx(GT)")  table(data$"NO2(GT)")  table(data$T)  table(data$RH)  table(data$AH)  # 选择变量并计算它们的相关系数矩阵  correlation\_matrix <- cor(data[c("CO(GT)", "NMHC(GT)", "C6H6(GT)","NOx(GT)","NO2(GT)")], use = "complete.obs", method = "pearson")  print(correlation\_matrix)  # cor.test相关性检验  var1 <- data$`CO(GT)`  var2 <- data$`NMHC(GT)`  # 使用 cor.test 进行皮尔逊相关性检验  test\_result <- cor.test(var1, var2, alternative = "two.sided", method = "pearson")  print(test\_result)  # cor2pcor计算偏相关系数矩阵  if (!require(corpcor)) {  install.packages("corpcor")}  library(corpcor)  vars <- data[, c("CO(GT)", "NMHC(GT)", "NO2(GT)")]  # 计算相关系数矩阵  cor\_matrix <- cor(vars, use = "complete.obs")  # 使用 cor2pcor 函数将相关系数矩阵转换为偏相关系数矩阵  pcor\_matrix <- cor2pcor(cor\_matrix)  print(pcor\_matrix)  #卡方检验  RH\_clean <- na.omit(data$RH)  AH\_clean <- na.omit(data$AH)  # 将 RH 和 AH 离散化为 3 个区间  RH\_binned <- cut(RH\_clean, breaks = 3, labels = c("Low", "Medium", "High"))  AH\_binned <- cut(AH\_clean, breaks = 3, labels = c("Low", "Medium", "High"))  # 创建 RH 和 AH 的列联表  contingency\_table\_RH\_AH <- table(RH\_binned, AH\_binned)  # 进行卡方检验  chisq\_test\_result\_RH\_AH\_no\_correction <- chisq.test(contingency\_table\_RH\_AH, correct = FALSE)  print(chisq\_test\_result\_RH\_AH\_no\_correction)  #计算边际频数 边际百分比  # 创建列联表  contingency\_table\_RH\_AH <- table(RH\_binned, AH\_binned)  # 计算 RH（相对湿度）的边际频数  marginal\_RH <- margin.table(contingency\_table\_RH\_AH, margin = 1) # margin = 1 表示按行计算边际频数  # 计算 AH（绝对湿度）的边际频数  marginal\_AH <- margin.table(contingency\_table\_RH\_AH, margin = 2) # margin = 2 表示按列计算边际频数  print("边际频数（RH）:")  print(marginal\_RH)  print("边际频数（AH）:")  print(marginal\_AH)  # 计算 RH（相对湿度）的边际百分比  prop\_RH <- prop.table(contingency\_table\_RH\_AH, margin = 1)\*100 # margin = 1 按行计算比例  # 计算 AH（绝对湿度）的边际百分比  prop\_AH <- prop.table(contingency\_table\_RH\_AH, margin = 2) \*100 # margin = 2 按列计算比例  # 输出边际百分比  print(prop\_RH)  print(prop\_AH)  #茎叶图  stem(data$T)  #箱线图  variables <- c("CO(GT)", "NMHC(GT)", "C6H6(GT)", "NOx(GT)", "NO2(GT)", "T", "RH", "AH")  variable\_names <- c("CO(GT)" = "一氧化碳浓度",  "NMHC(GT)" = "非甲烷总烃浓度",  "C6H6(GT)" = "苯浓度",  "NOx(GT)" = "氮氧化物浓度",  "NO2(GT)" = "二氧化氮浓度",  "T" = "温度",  "RH" = "相对湿度",  "AH" = "绝对湿度")  par(mfrow = c(2, 4), mar = c(5, 5, 3, 1))  for (var in variables) {  boxplot(data[[var]],  main = paste(variable\_names[var], "的箱线图"),  ylab = variable\_names[var],  col = "lightblue",  border = "darkblue",  notch = TRUE,  outline = TRUE,  las = 1,  cex.axis = 1.2,  cex.lab = 1.2)  }  #直方图和密度图  T\_clean <- na.omit(data$T)  RH\_clean <- na.omit(data$RH)  AH\_clean <- na.omit(data$AH)  par(mfrow = c(3, 2))  # T 的直方图和密度图  hist(T\_clean, main = "T 的直方图", xlab = "T (温度)", col = "skyblue", border = "white")  plot(density(T\_clean), main = "T 的密度图", xlab = "T (温度)", col = "blue", lwd = 2)  # RH 的直方图和密度图  hist(RH\_clean, main = "RH 的直方图", xlab = "RH (相对湿度)", col = "lightgreen", border = "white")  plot(density(RH\_clean), main = "RH 的密度图", xlab = "RH (相对湿度)", col = "darkgreen", lwd = 2)  # AH 的直方图和密度图  hist(AH\_clean, main = "AH 的直方图", xlab = "AH (绝对湿度)", col = "orange", border = "white")  plot(density(AH\_clean), main = "AH 的密度图", xlab = "AH (绝对湿度)", col = "red", lwd = 2)  #绘制小提琴图  if (!require(vioplot)) {  install.packages("vioplot")  }  library(vioplot)  T\_clean <- na.omit(data$T)  RH\_clean <- na.omit(data$RH)  AH\_clean <- na.omit(data$AH)  dev.new(width = 10, height = 8)  par(mar = c(5, 5, 4, 2) + 0.1)  # 绘制小提琴图  vioplot(T\_clean, RH\_clean, AH\_clean,  names = c("T (温度)", "RH (相对湿度)", "AH (绝对湿度)"),  col = c("skyblue", "lightgreen", "orange"))  title("T、RH 和 AH 的小提琴图", cex.main = 1.5) # 调整标题大小  #克利夫兰图  T\_clean <- na.omit(data$T)  RH\_clean <- na.omit(data$RH)  AH\_clean <- na.omit(data$AH)  par(mfrow = c(1, 3))  # 绘制 T 的克利夫兰图  dotchart(T\_clean[1:100], main = "T 的克利夫兰图", xlab = "T (温度)", col = "blue")  # 绘制 RH 的克利夫兰图  dotchart(RH\_clean[1:100], main = "RH 的克利夫兰图", xlab = "RH (相对湿度)", col = "green")  # 绘制 AH 的克利夫兰图  dotchart(AH\_clean[1:100], main = "AH 的克利夫兰图", xlab = "AH (绝对湿度)", col = "orange")  par(mfrow = c(1, 1))  #绘制折线图  T\_clean <- na.omit(data$T)  RH\_clean <- na.omit(data$RH)  AH\_clean <- na.omit(data$AH)  # 检查三者是否长度一致（截取最小长度）  min\_length <- min(length(T\_clean), length(RH\_clean), length(AH\_clean))  T\_clean <- T\_clean[1:min\_length]  RH\_clean <- RH\_clean[1:min\_length]  AH\_clean <- AH\_clean[1:min\_length]  # 绘制 T 的线图  plot(T\_clean, type = "l", col = "blue", lwd = 2, ylim = range(c(T\_clean, RH\_clean, AH\_clean)),  xlab = "索引", ylab = "数值", main = "T、RH 和 AH 的线图",  cex.main = 1.5, cex.lab = 1.2, cex.axis = 1.1)  # 添加 RH 的线图  lines(RH\_clean, col = "green", lwd = 2, lty = 2)  # 添加 AH 的线图  lines(AH\_clean, col = "orange", lwd = 2, lty = 3)  # 添加网格线  grid(lwd = 1, col = "gray", lty = "dotted") # 添加灰色虚线网格  # 添加图例  legend("topright", legend = c("T (温度)", "RH (相对湿度)", "AH (绝对湿度)"),  col = c("blue", "green", "orange"), lty = c(1, 2, 3), lwd = 2,  cex = 1.2, box.lwd = 2, bg = "white")  # 绘制柱状图  CO\_clean <- as.numeric(na.omit(data$`CO(GT)`))  barplot(CO\_clean,  main = "CO(GT) 的柱状图",  ylab = "CO (GT)",  col = rainbow(length(CO\_clean)),  border = "black",  space = 0.5,  las = 2)  # 绘制饼图  CO\_clean <- as.numeric(na.omit(data$`CO(GT)`))  # 划定 CO(GT) 的区间  breaks <- c(0, 1, 3, 6, 12)  CO\_groups <- cut(CO\_clean, breaks = breaks, right = FALSE, labels = c("0-1", "1-3", "3-6", "6-12"))  # 计算每个区间的频数  CO\_freq <- table(CO\_groups)  pie(CO\_freq,  main = "CO(GT) 的区间分布",  col = rainbow(length(CO\_freq)),  labels = paste(names(CO\_freq), "\n", CO\_freq),  cex.main = 1.5,  cex.lab = 1.2)  # 使用 plot 绘制散点图  plot(RH\_clean,  type = "p",  col = rgb(0, 1, 0, 0.5),  pch = 19,  ylab = "RH (相对湿度)",  main = "RH 散点图",  cex.main = 1.6,  cex.lab = 1.3,  cex.axis = 1.2,  col.axis = "darkblue",  col.lab = "darkgreen",  las = 1,  bg = "lightgray")  # 使用 smoothScatter 绘制平滑散点图  RH\_clean <- na.omit(data$RH)  # 使用 smoothScatter 绘制平滑散点图  smoothScatter(1:length(RH\_clean), RH\_clean,  main = "RH的高密度散点图",  xlab = "Index",  ylab = "RH (Relative Humidity)",  colramp = colorRampPalette(c("white", "blue", "green", "yellow", "red")), # 设置渐变色  pch = 16)  #使用 pairs 绘制矩阵散点图，并添加回归线  #载入包  if (!requireNamespace("car", quietly = TRUE)) {  install.packages("car")  }  data\_matrix <- data.frame(T\_clean, RH\_clean, AH\_clean)  panel\_with\_regression <- function(x, y, ...) {  # 绘制散点图  points(x, y, ...)  # 绘制回归线  abline(lm(y ~ x), col = "red", lwd = 2) # 红色回归线  }  # 使用 pairs 绘制矩阵散点图，并添加回归线  pairs(data\_matrix,  main = "散点图矩阵带回归线",  col = rgb(0, 0, 1, 0.5),  pch = 19,  cex.main = 1.6,  cex.lab = 1.3,  cex.axis = 1.2,  col.axis = "darkblue",  col.lab = "darkgreen",  panel = panel\_with\_regression)  # 绘制 Q-Q 图  # 设置图形布局：1行3列  par(mfrow = c(1, 3))  # 绘制 T\_clean (温度) 的 Q-Q 图  qqnorm(T\_clean, main = "温度 (T) 的 Q-Q 图", col = "blue", pch = 19)  qqline(T\_clean, col = "red", lwd = 2)  # 绘制 AH\_clean (绝对湿度) 的 Q-Q 图  qqnorm(AH\_clean, main = "绝对湿度 (AH) 的 Q-Q 图", col = "green", pch = 19)  qqline(AH\_clean, col = "orange", lwd = 2)  # 绘制 RH\_clean (相对湿度) 的 Q-Q 图  qqnorm(RH\_clean, main = "相对湿度 (RH) 的 Q-Q 图", col = "purple", pch = 19)  qqline(RH\_clean, col = "brown", lwd = 2)  #检验方差同质性  if (!requireNamespace("car", quietly = TRUE)) {  install.packages("car")  }  if (!requireNamespace("pwr", quietly = TRUE)) {  install.packages("pwr")  }  library(car)  library(pwr)  # 提取 6:00 和 18:00 的数据，并去除 CO(GT) 列的 NA 值  data\_combined <- na.omit(data.frame(  CO\_GT = c(data$`CO(GT)`[data$Time == "6:00:00"],  data$`CO(GT)`[data$Time == "18:00:00"]),  Time = factor(rep(c("6:00 AM", "6:00 PM"),  times = c(sum(data$Time == "6:00:00"), sum(data$Time == "18:00:00"))))  ))  # Levene's Test 检验方差同质性  levene\_result <- leveneTest(CO\_GT ~ Time, data = data\_combined)  print(levene\_result)  # 独立样本 t 检验  data$Time <- as.character(data$Time)  data\_6am <- na.omit(data$`CO(GT)`[data$Time == "6:00:00"]) # 6:00 的数据  data\_6pm <- na.omit(data$`CO(GT)`[data$Time == "18:00:00"]) # 18:00 的数据  # 独立样本 t 检验（假设两组数据的方差不等）  t\_test\_result <- t.test(  data\_6am, data\_6pm,  paired = FALSE, # 独立样本  var.equal = FALSE, # 假定方差不等  mu = 0, # 假设均值差为 0  alternative = "two.sided" # 双尾检验  )  print(t\_test\_result)  # 计算描述性统计量进行功效分析  mean\_6am <- mean(data\_6am)  mean\_6pm <- mean(data\_6pm)  sd\_6am <- sd(data\_6am)  sd\_6pm <- sd(data\_6pm)  n\_6am <- length(data\_6am)  n\_6pm <- length(data\_6pm)  # 计算效应量 (Cohen's d)  pooled\_sd <- sqrt(((n\_6am - 1) \* sd\_6am^2 + (n\_6pm - 1) \* sd\_6pm^2) / (n\_6am + n\_6pm - 2))  d <- (mean\_6am - mean\_6pm) / pooled\_sd # Cohen's d  library(pwr)  power\_analysis <- pwr.t2n.test(  d = d, # 效应量  n1 = n\_6am, # 第一组样本量  n2 = n\_6pm, # 第二组样本量  sig.level = 0.05, # 显著性水平  power = NULL, # 计算功效  alternative = "two.sided" # 双尾检验  )  print(power\_analysis)  #Wilcoxon 秩和检验  # 使用 Wilcoxon 秩和检验比较 6:00 AM 和 6:00 PM 的数据  wilcox\_result <- wilcox.test(  data\_6am, data\_6pm,  paired = FALSE, # 两组是独立样本  alternative = "two.sided" # 双尾检验  )  print("Wilcoxon Test Results for 6:00 AM vs 6:00 PM:")  print(wilcox\_result)  # 进行置换检验  if (!requireNamespace("coin", quietly = TRUE)) {  install.packages("coin")  }  library(coin)  # 组合 6:00 AM 和 6:00 PM 的数据  data\_combined <- data.frame(  CO\_GT = c(data\_6am, data\_6pm), # `CO(GT)` 浓度  Time = factor(rep(c("6:00 AM", "6:00 PM"),  times = c(length(data\_6am), length(data\_6pm)))) # 时间变量  )  perm\_test\_result <- oneway\_test(CO\_GT ~ Time, data = data\_combined, distribution = "approximate")  print("Permutation Test Results for 6:00 AM vs 6:00 PM:")  print(perm\_test\_result)  #单因素方差分析  anova\_result <- aov(`CO(GT)` ~ AH, data )  anova(anova\_result)  # 去除包含 NA 的行，并确保数据框长度一致  data\_clean <- data[!is.na(data[["CO(GT)"]]) & !is.na(data[["AH"]]), ]  # 重新创建清洗后的 CO 和 AH 列  CO\_clean <- data\_clean[["CO(GT)"]]  AH\_clean <- data\_clean[["AH"]]  # 添加清洗后的 CO 和 AH 列  data\_clean$CO\_clean <- CO\_clean  data\_clean$AH\_clean <- AH\_clean  # 添加分组变量  data\_clean$group <- ifelse(data\_clean$Time >= "18:00:00" & data\_clean$Time <= "22:00:00", "实验组", "对照组")  ancova\_model <- aov(CO\_clean ~ AH\_clean + group, data = data\_clean)  # 显示结果  summary(ancova\_model)  # 多因素方差分析  data\_clean <- na.omit(data)  # 多因素方差分析模型，查看T, RH, AH对CO(GT)的影响  model <- aov(`NO2(GT)` ~ T + RH + AH, data = data\_clean)  # 显示结果  anova(model)  #线性回归模型  #创建一个新的数据框 processed\_data  processed\_data <- data  # 将 Date 列转换为 Date 类型，并提取 Month  processed\_data$Date <- as.Date(processed\_data$Date, format = "%Y/%m/%d")  processed\_data$Month <- as.numeric(format(processed\_data$Date, "%m"))  # 将 Time 列转换为 POSIXct 类型，并提取 Hour  processed\_data$Time <- strptime(processed\_data$Time, format = "%H:%M:%S")  processed\_data$Hour <- as.numeric(format(processed\_data$Time, "%H"))  # 线性回归模型：NO2(GT) ~ T + AH + Month + Hour  model <- lm(`NO2(GT)` ~ T + AH + Month + Hour, data = processed\_data)  # 查看模型系数  print(coefficients(model))  # 查看模型摘要  print(summary(model))  # 获取模型系数的置信区间（95% 置信区间）  conf\_intervals <- confint(model, level = 0.95)  print(conf\_intervals)  new\_data <- data.frame(  T = c(20.5, 22.3, 18.2), # 新的温度数据  AH = c(0.80, 0.75, 0.78), # 新的绝对湿度数据  Date = as.Date(c("2024/11/27", "2024/11/27", "2024/11/27")), # 新的日期  Time = c("14:30:00", "15:00:00", "16:30:00") # 新的时间  )  # 将新数据中的 Date 和 Time 转换为 Month 和 Hour  new\_data$Month <- as.numeric(format(new\_data$Date, "%m"))  new\_data$Hour <- as.numeric(format(strptime(new\_data$Time, format = "%H:%M:%S"), "%H"))  # 使用 predict() 函数生成新数据的预测值  predictions <- predict(model, newdata = new\_data, type = "response")  # 查看预测结果  print(predictions)  #模型参数  hatvalues(model)  rstudent(model)  cooks.distance(model)  library(car)  vif(model) |
| **其他材料与附件** | **说明：放一个数据集的截图，然后附上数据集作为独立附件文件** |