# 茂名市第一水质净化厂COD浓度分析报告

**项目摘要：**本报告旨在分析茂名市第一水质净化厂污水进厂COD浓度在8月11日前后的动态变化情况。为了评估这一变化，我们采用了多种数据分析方法，包括数据清洗与降噪、统计分析方法如概率密度分布和非参数统计检验（Mann-Whitney U 测试）等。在数据清洗与降噪方面，我们致力于确保数据的准确性和可信度，去除了重复数据和异常值。接着，我们采用了概率密度分布来观察COD浓度的分布情况，以及Mann-Whitney U 测试来比较8月11日前后的数据样本是否存在显著差异。最终的分析结果显示，在8月11日后，污水进厂COD浓度发生了明显的变化，呈下降趋势。这一发现对水质管理和管网维护具有重要意义，有助于满足一污方面的要求，同时也为未来进一步的水质监测和管网维护提供了科学支持。

1. **项目背景**

茂名市第一水质净化厂于8月11日接入新的污水管网。为满足一污方面的要求，需要验证污水的历史COD浓度是否在8月11日前后发生了显著变化。

1. **分析方法**

本项目采用了以下数据分析方法，这些方法有助于准确地评估茂名市第一水质净化厂的污水进厂COD浓度变化情况，并为结果提供有力支撑。

1. 数据清洗与降噪方法：

- 去重（Duplicate Removal）：我们首先对原始数据进行去重处理，以消除重复采集的数据点。这有助于减少重复数据可能带来的干扰。

- 异常值处理（Outlier Handling）：我们采用了标准差方法（Standard Deviation Method）对数据进行异常值检测和处理，以识别和排除可能由于测量或采集错误引起的异常值。这有助于提高数据的准确性和可信度。

1. 统计分析方法：

- 概率密度分布（Probability Density Distribution）：我们采用概率密度分布来观察COD浓度的分布情况。这有助于了解COD浓度在不同时间段内的分布特征，是否存在显著的变化。

- Mann-Whitney U 测试：我们使用Mann-Whitney U 测试来比较8月11日前后的数据样本是否存在显著差异。这是一种非参数统计检验方法，适用于不满足正态分布假设的数据。

- Student's t test 测试：我们假设8.11前后数据近似呈现正态分布，使用t检验观察两者差异是否明显以及均值大小的关系。

选择理由：概率密度分布允许我们可视化COD浓度的变化趋势，而Mann-Whitney U 测试是一种鲁棒的方法，不依赖于数据分布的假设，适用于非正态分布的数据。同时 Student's t test 测试可以在假设条件下观察数据差异是否明显以及均值大小的关系。这些方法的选择基于数据的性质和我们需要验证的问题，有助于全面评估COD浓度变化。

通过这些数据清洗与降噪方法和统计分析方法的综合应用，我们可以准确地分析8月11日前后的COD浓度数据，为研究结果提供可信度和科学支持。

1. **分析结果**

通过采用严格的数据清洗、降噪以及统计分析方法，我们得出了以下关于茂名市第一水质净化厂污水进厂COD浓度变化的分析结果。

1. 数据清洗与降噪：

首先我们对原始数据进行了统计分析，先统计每天的数据量，观察分布是否平均：

|  |  |  |
| --- | --- | --- |
| date | COD\_original | COD\_unique |
| 2023/7/11 | 986 | 13 |
| 2023/7/12 | 1062 | 13 |
| 2023/7/13 | 1008 | 14 |
| 2023/7/14 | 965 | 13 |
| 2023/7/15 | 1014 | 11 |
| 2023/7/16 | 1061 | 12 |
| 2023/7/17 | 622 | 5 |
| 2023/7/18 | 115 | 9 |
| 2023/7/19 | 106 | 11 |
| 2023/7/20 | 46 | 12 |
| 2023/7/21 | 102 | 13 |
| 2023/7/22 | 57 | 13 |
| 2023/7/23 | 122 | 14 |
| 2023/7/24 | 136 | 12 |
| 2023/7/25 | 90 | 6 |
| 2023/7/26 | 2 | 1 |
| 2023/8/2 | 91 | 5 |
| 2023/8/3 | 14 | 3 |
| 2023/8/4 | 1 | 1 |
| 2023/8/5 | 56 | 14 |
| 2023/8/6 | 47 | 12 |
| 2023/8/7 | 59 | 14 |
| 2023/8/8 | 85 | 13 |
| 2023/8/9 | 141 | 14 |
| 2023/8/10 | 51 | 12 |
| 2023/8/11 | 49 | 12 |
| 2023/8/12 | 79 | 11 |
| 2023/8/13 | 47 | 12 |
| 2023/8/14 | 86 | 12 |
| 2023/8/15 | 25 | 6 |
| 2023/8/16 | 12 | 1 |
| 2023/8/17 | 33 | 9 |
| 2023/8/18 | 69 | 13 |
| 2023/8/19 | 26 | 4 |
| 2023/8/20 | 6 | 3 |
| 2023/8/23 | 97 | 9 |
| 2023/8/24 | 7 | 3 |
| 2023/8/25 | 32 | 5 |

表 1 去重前后每日数据量统计

经统计分析，8月11日之后的数据每天不足100条，去掉重复发送的COD数据后每天的数据只剩10条左右，并且在每天不足100条的数据中，还有不少重复发送的量。为了探究重复数据出现原因，计算了每段连续相同COD数据的持续时间，然后观察每天出现多少段连续COD情况。

|  |  |
| --- | --- |
| Duration Interval | Count |
| 1-2小时 | 6462 |
| 4-5小时 | 564 |
| 5小时及以上 | 518 |
| 3-4小时 | 493 |
| 0-1小时 | 375 |
| 2-3小时 | 186 |

表 2 连续相同COD持续时间及数量统计

通过观察连续相同COD的持续时间，不同持续区间的数量，可见大部分COD持续了1-2小时，可以假设持续2小时以上的COD数值不变的采集数据可能为问题数据，为了数据的稳定性，决定采取去重策略。

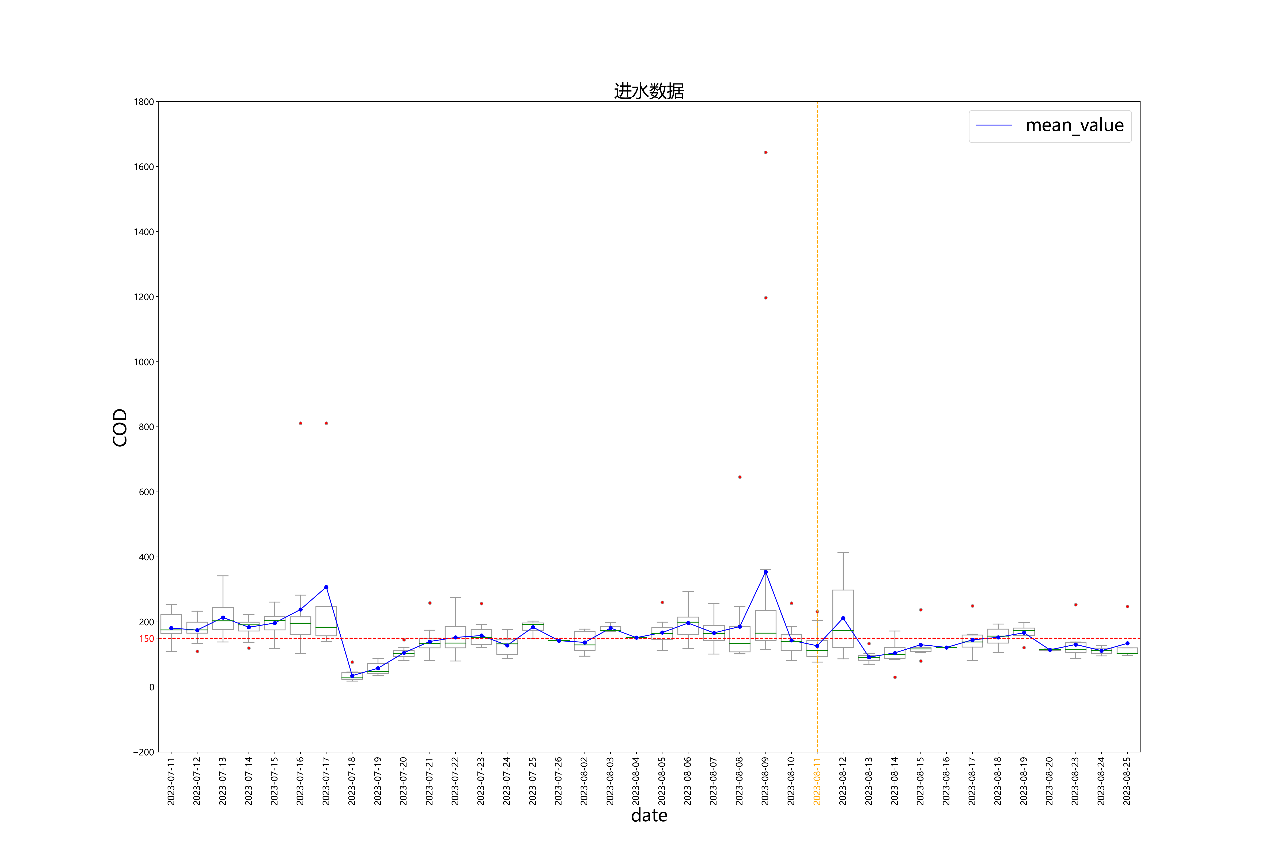


图 1 去重后按日统计箱型图

异常值处理：

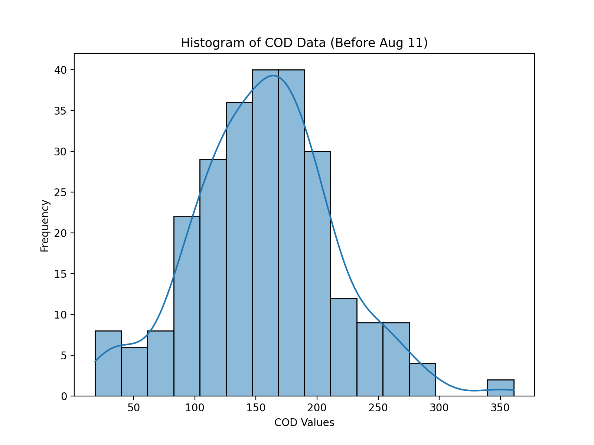
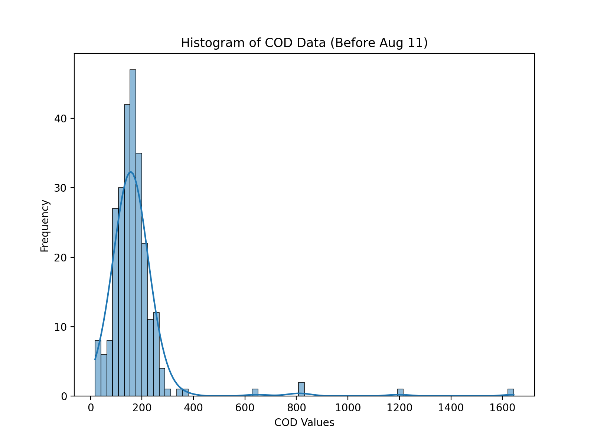


图 2 8月11日前异常值处理前后数据分布情况

在观察11日前后数据是否符合正态分布时发现，异常值影响了数据分布的分析，故采用了标准差方法（Standard Deviation Method）进行异常值处理。

|  |  |
| --- | --- |
| 异常值数量 | 5 |
| 异常值所占比例 | 1.39% |

表3异常值统计

经过对原始数据进行的去重和异常值处理，确保了数据的准确性和稳定性。这一步骤是为了消除重复数据和排除可能的测量或采集错误。

1. 时段大小变化分析:

每天各小时数据量的频数图显示如下，每天传输的数据量没有按小时均匀分布的特点。

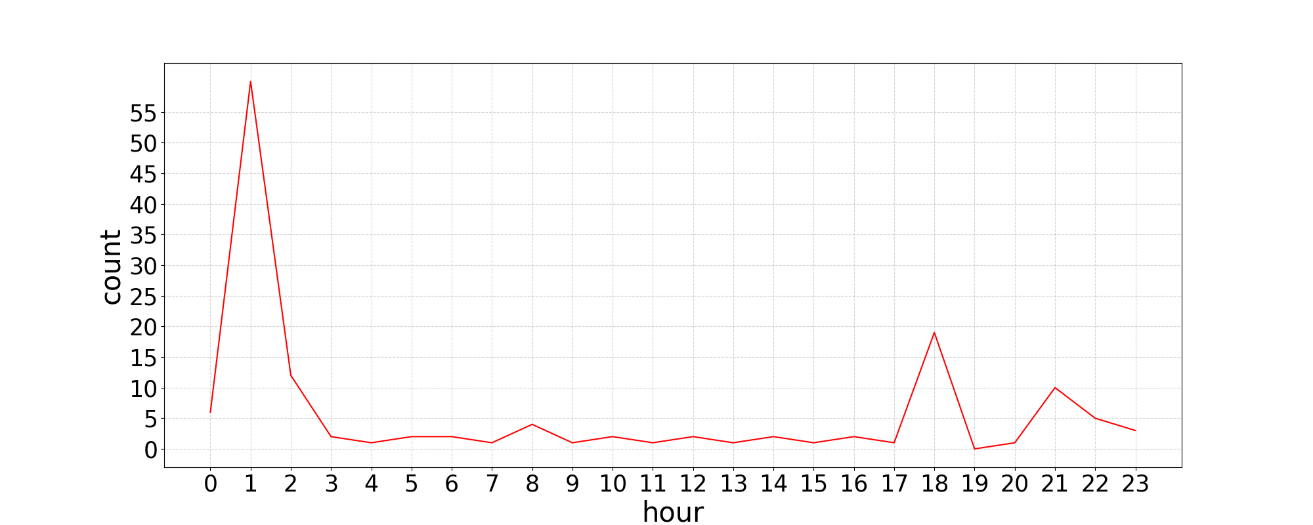


图 3 8.9 24小时数据量统计折线图

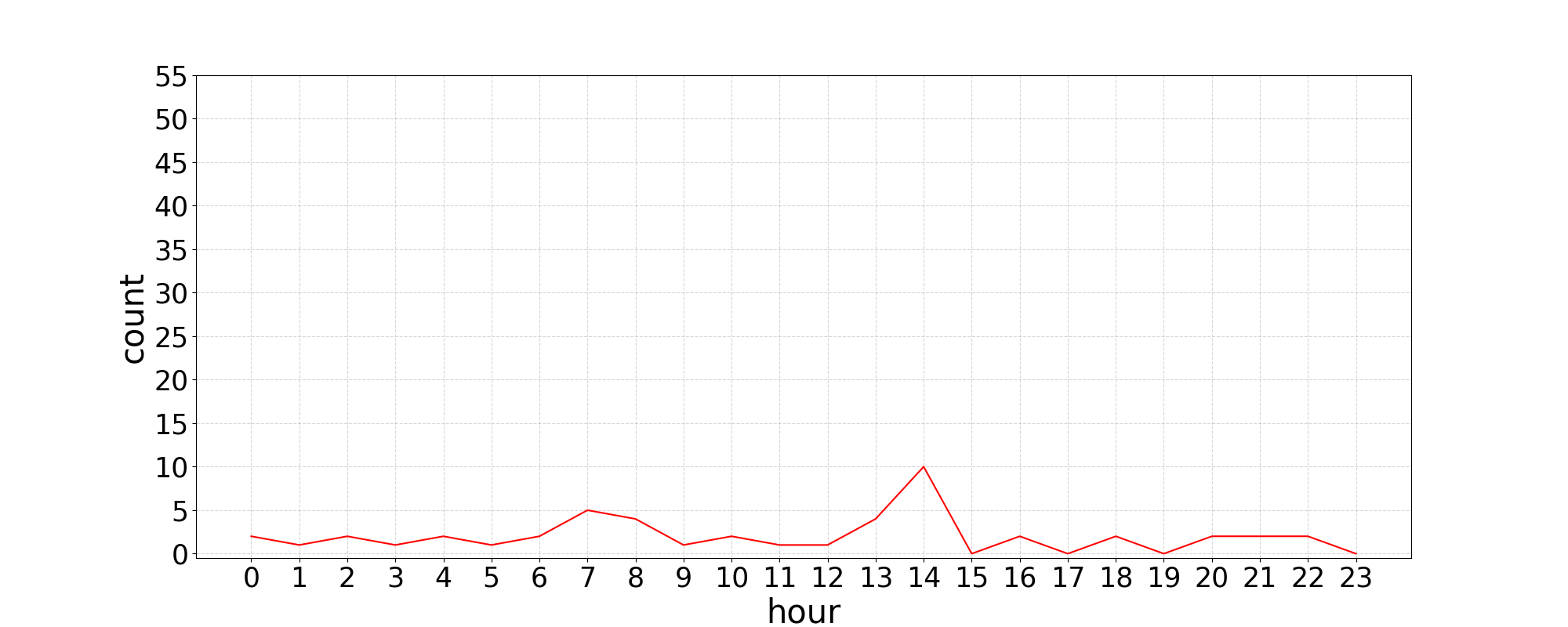


图 4 8.11 24小时数据量统计折线图

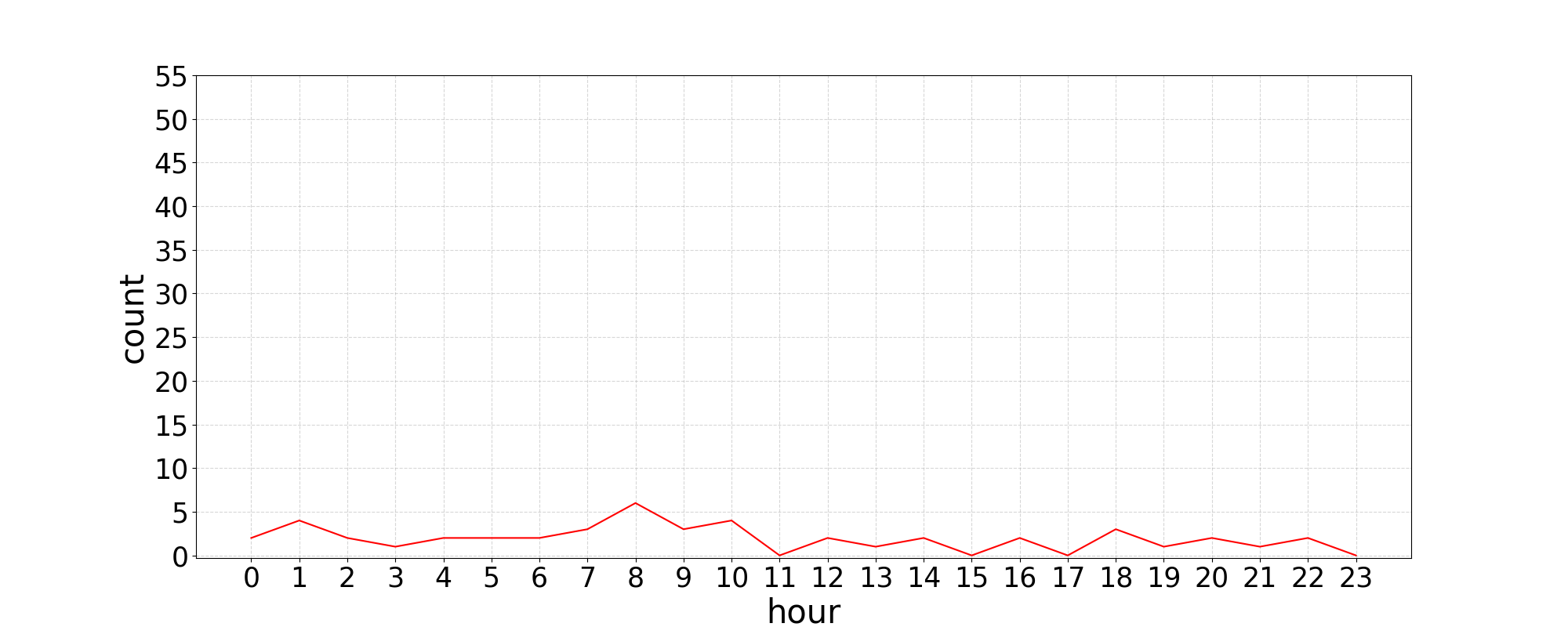


图 5 8.13 24小时数据量统计折线图

8.11前后数据量统计显示：未去重前，8.11前后24小时各小时数据量差异较大，去重后差异不大，但数据量过少，没有足够的数据支撑，故无法讨论时段大小变化。

|  |  |  |
| --- | --- | --- |
| 小时 | 8.11日前各小时数据量 | 8.11日后各小时数据量 |
| 0 | 346 | 16 |
| 1 | 364 | 13 |
| 2 | 313 | 18 |
| 3 | 322 | 14 |
| 4 | 354 | 21 |
| 5 | 343 | 30 |
| 6 | 349 | 30 |
| 7 | 349 | 17 |
| 8 | 354 | 17 |
| 9 | 392 | 20 |
| 10 | 359 | 24 |
| 11 | 314 | 26 |
| 12 | 322 | 52 |
| 13 | 373 | 26 |
| 14 | 353 | 57 |
| 15 | 330 | 14 |
| 16 | 318 | 16 |
| 17 | 311 | 7 |
| 18 | 329 | 20 |
| 19 | 278 | 32 |
| 20 | 333 | 24 |
| 21 | 329 | 41 |
| 22 | 322 | 26 |
| 23 | 282 | 7 |

表 4 未去重前各小时数据量

|  |  |  |
| --- | --- | --- |
| 小时 | 8.11日前各小时数据量 | 8.11日后各小时数据量 |
| 0 | 18 | 7 |
| 1 | 0 | 0 |
| 2 | 21 | 8 |
| 3 | 0 | 1 |
| 4 | 18 | 7 |
| 5 | 2 | 2 |
| 6 | 18 | 8 |
| 7 | 0 | 1 |
| 8 | 19 | 4 |
| 9 | 2 | 0 |
| 10 | 20 | 5 |
| 11 | 1 | 0 |
| 12 | 13 | 6 |
| 13 | 1 | 1 |
| 14 | 21 | 7 |
| 15 | 0 | 0 |
| 16 | 19 | 7 |
| 17 | 0 | 0 |
| 18 | 21 | 8 |
| 19 | 1 | 1 |
| 20 | 15 | 6 |
| 21 | 9 | 4 |
| 22 | 24 | 12 |
| 23 | 17 | 5 |

表 5 去重后各小时数据量

1. 概率密度分布分析： 我们绘制了8月11日前后的COD数据量的概率密度分布图，用于观察COD数据量的分布情况。

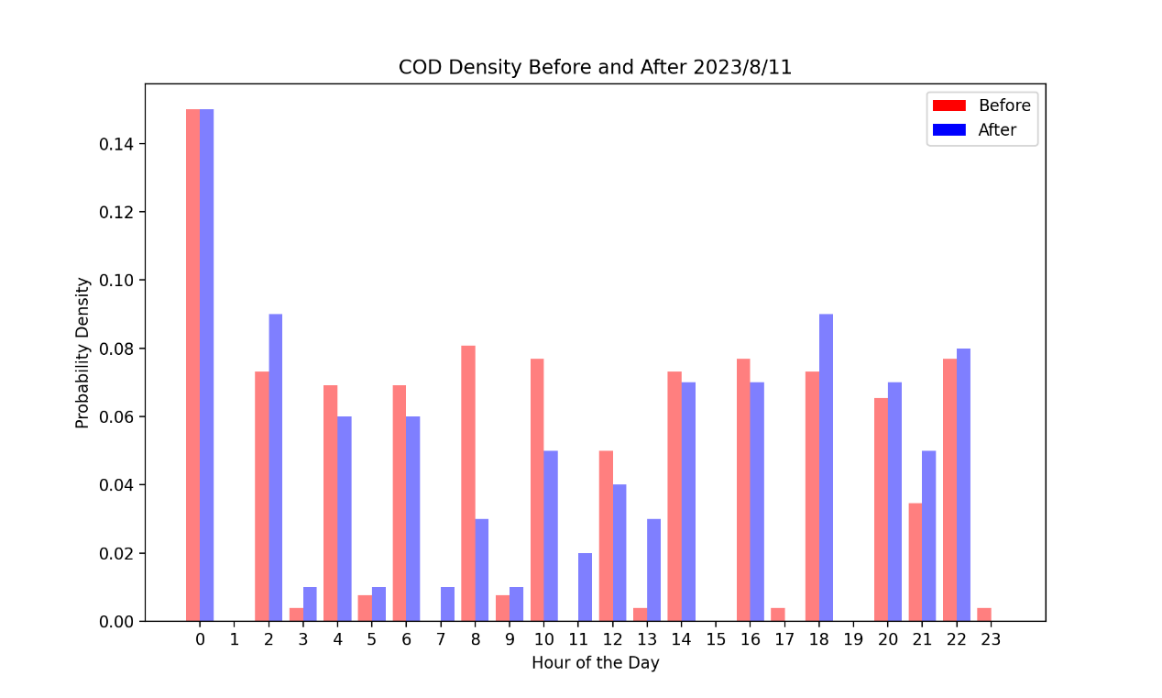


图6去重后概率密度分布

图表数据显示，8月11日前后的COD数据量分布相对匹配，故可分析整体变化。

1. Mann-Whitney U 测试：我们使用Mann-Whitney U 测试比较了8月11日前后的数据样本，以验证是否存在显著差异。

首先观察11日前后数据是否符合正态分布，以选择合适的差异性检测方法

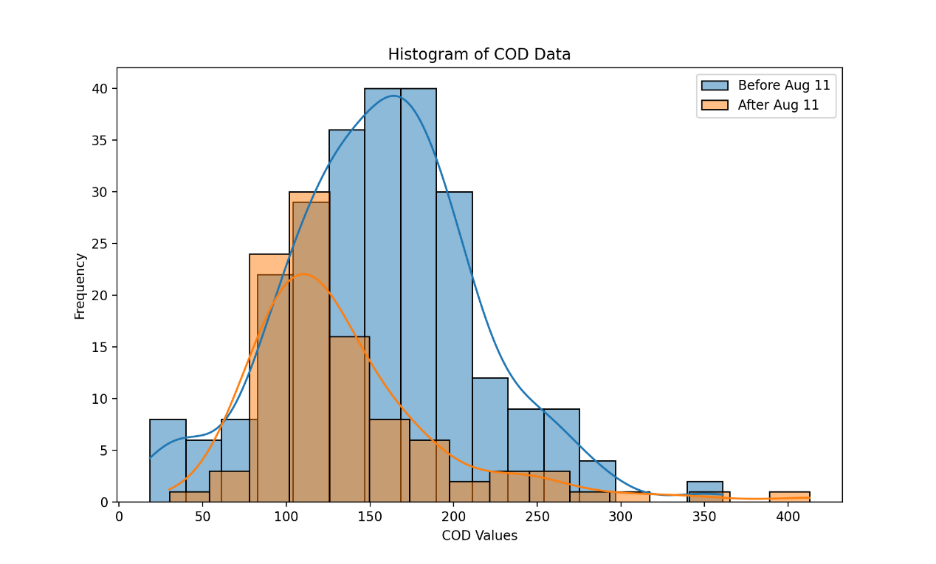


图 7 11日前后数据分布

经观察，11日前的数据趋向于正态分布，但是11日后的数据趋向于偏态分布，故决定采取Mann-Whitney U 测试方法来检测数据差异性，Mann-Whitney U 测试是一种非参数测试方法，不受数据分布的影响，因此适用于比较不同分布的数据。

具体来说，Mann-Whitney U 测试的目标是确定两组数据中的中位数是否存在显著差异。它基于样本中的秩次来执行比较。测试的零假设是两组数据的中位数相等，备择假设是两组数据的中位数不相等。如果测试结果表明p值小于显著性水平，则可得出结论，两组数据的中位数存在显著差异。这使得 Mann-Whitney U 测试在不满足正态分布假设或需要比较中位数的情况下非常有用。在计算过程中设置显著性水平alpha = 0.05。

|  |  |
| --- | --- |
| Mann-Whitney U 测试结果 | 8月11日前后的数据在中位数上存在显著差异 |
| Mann-Whitney U P 值 | 0.00000616 |

表 6 Mann-Whitney U 测试结果

Mann-Whitney U 测试的结果显示，8月11日前后的数据存在显著差异，拒绝了两者数据分布相同的原假设。这进一步支持了COD浓度发生了变化的观点。

1. Student's t test 测试：假设8.11前后数据近似呈现正态分布，通过t检验观察两者差异是否明显以及均值大小的关系。

假设8.11前后数据近似呈现正态分布，采用Student's t test 测试执行独立双样本比较两组8.11前后样本之间的均值差异，通常用于检查两组数据是否具有显著的统计差异。

Student's t test 双边检验的零假设和备择假设分别为：H0: 样本均值 = 另一组样本均值 H1: 样本均值 ≠ 另一组样本均值。如果 P 值小于设定的显著性水平（通常为0.05），那么我们将拒绝零假设，认为两组数据在均值上存在显著差异。如果 P 值大于显著性水平，则没有足够的证据来拒绝零假设，表示两组数据在均值上没有显著差异。

通过t双边检验的结果，得到的P值为0.00353306小于设定的0.05可以得出两者差异显著，所以我们拒绝接受8.11前后数据呈正态分布且均值相等的零假设，前后数据有本质差异。同时通过检验得到的统计量t（样本均值之差与标准误的比值）为2.95757456大于0，可以得出8.11后的数据样本均值小于8.11前数据样本的均值的检验结果，这也为接入管道后COD浓度有所下降作支撑。

|  |  |
| --- | --- |
| t | P |
| 2.95757456 | 0.00353306 |

表 7 t双边检验的统计结果

1. 8.11前后六周箱型图对比图：观察8.11前四周和8.11后两周数据的箱型图，可以看出8.11接入管道后，8.11后传入数据的COD浓度较前四周有一定下降。

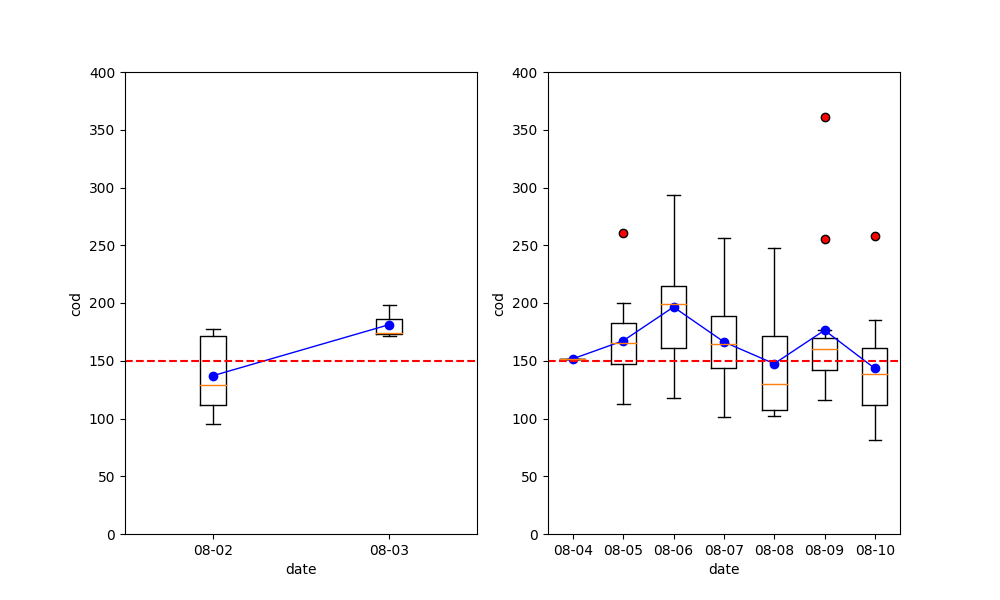
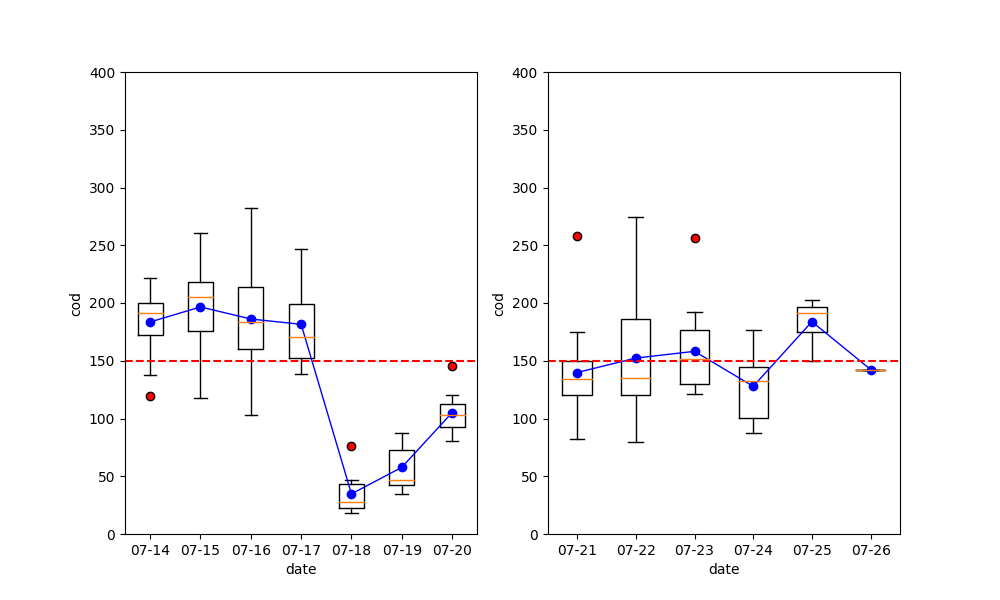


图 8 8.11前四周COD浓度箱型图

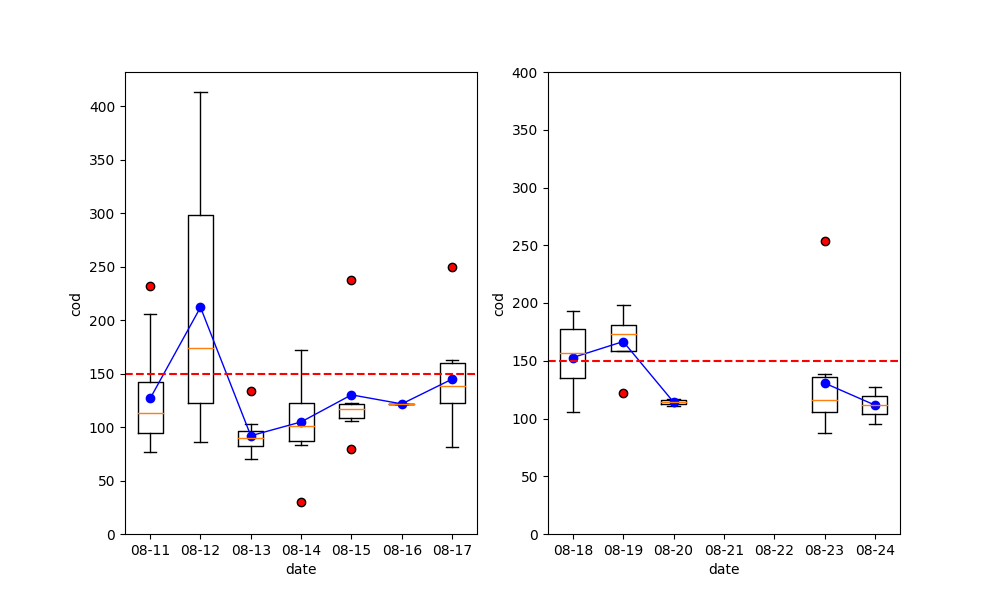


图 9 8.11后两周COD浓度箱型图

1. **总结**

通过数据分析，我们得出以下结论：

- 8月11日后，茂名市第一水质净化厂污水进厂COD浓度发生了明显下降。

- 去重和异常值处理有助于提高数据质量和分析结果的准确性。

- 概率密度分布和统计检验方法支持我们的结论，认为接入污水管对COD浓度的下降产生了正向作用。

- 时间序列分析表明，8月11日之后的数据质量较差，需要进一步改进数据采集方法。

这些结果有助于了解污水管道对水质的影响，为进一步的水质管理提供了有力支持。