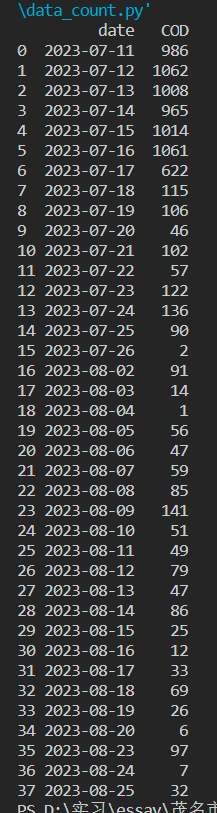
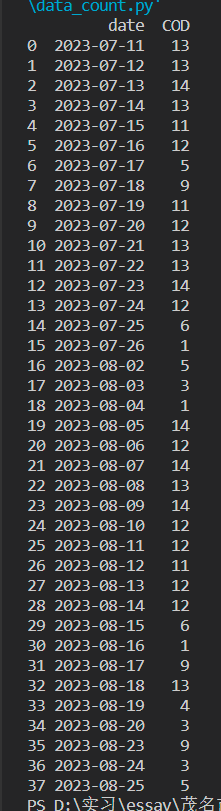
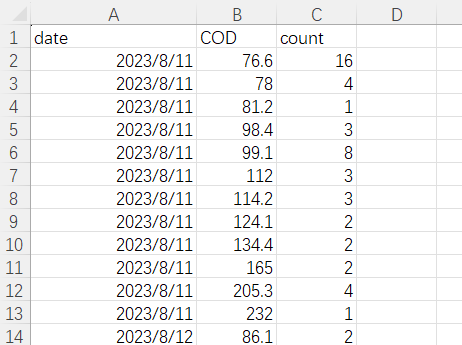
先对时间序列进行处理，单独分出天，小时，以便后续按照不同时间序列，对数据进行处理分析。

先统计每天的数据量，看分布是否平均

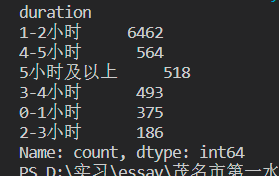
 

8月11日之后的数据每天不足100条，去掉重复发送的COD数据后每天的数据只剩10条左右

并且在每天不足100条的数据中，还有不少重复发送的量。

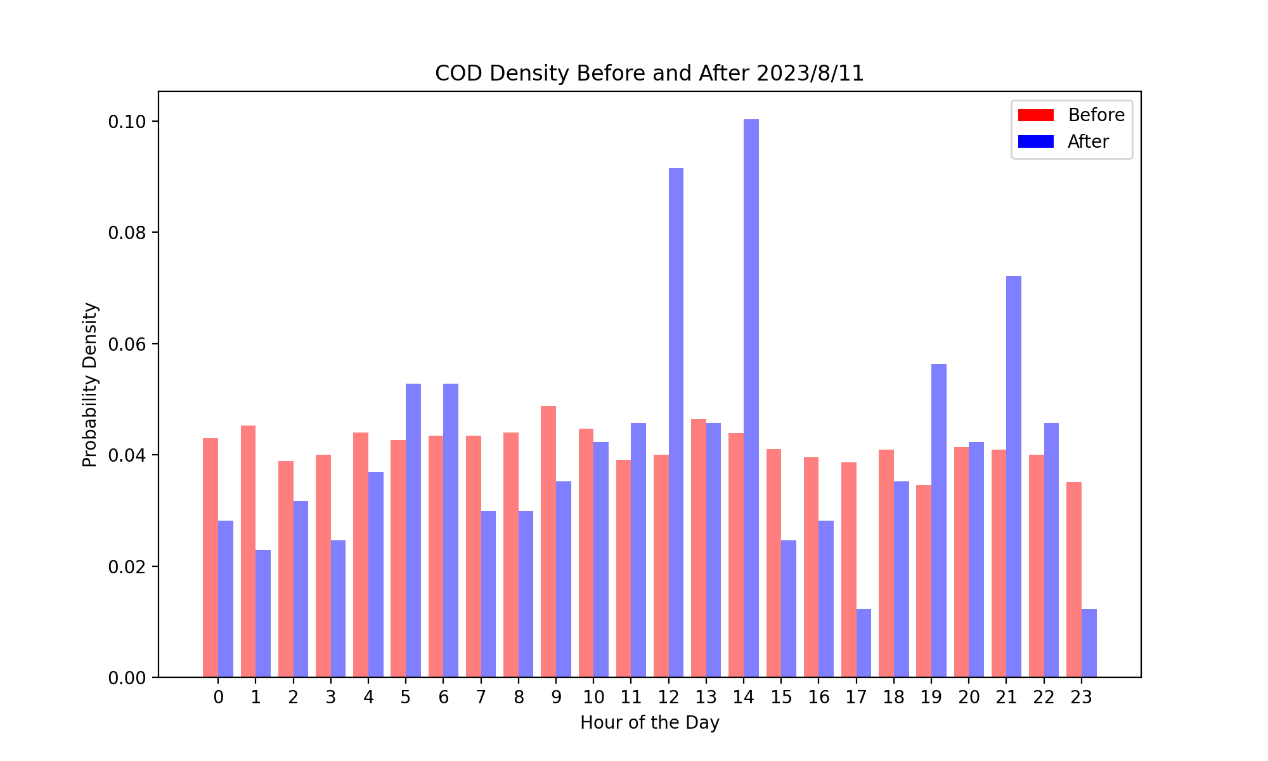


计算每段连续相同COD的持续时间，然后观察每天出现多少段连续COD情况



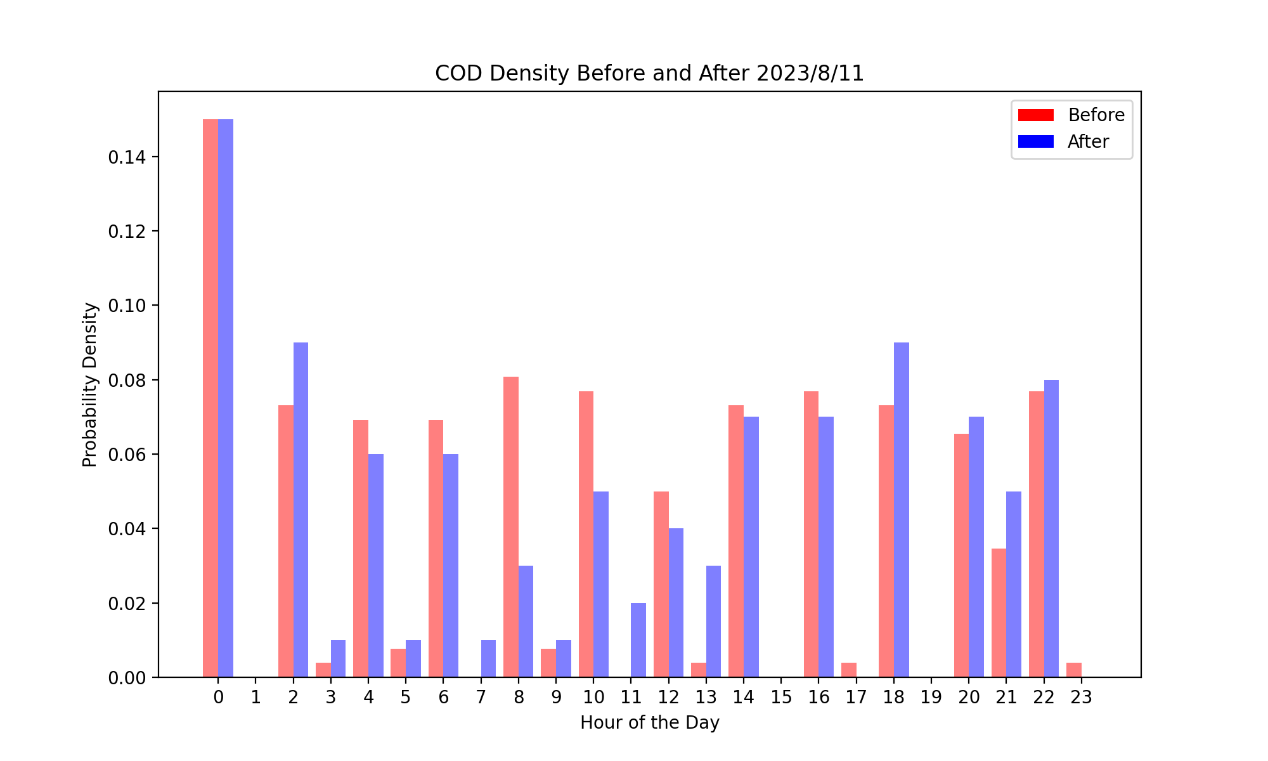
连续相同COD的持续时间，持续多久的数量，可见大部分COD持续了1-2小时，可以假设持续2小时以上的COD数值不变采集数据为问题数据，为了数据的稳定性，采取只保留0-2小时数据进行统计分析。

绘制24小时离散COD概率密度分布

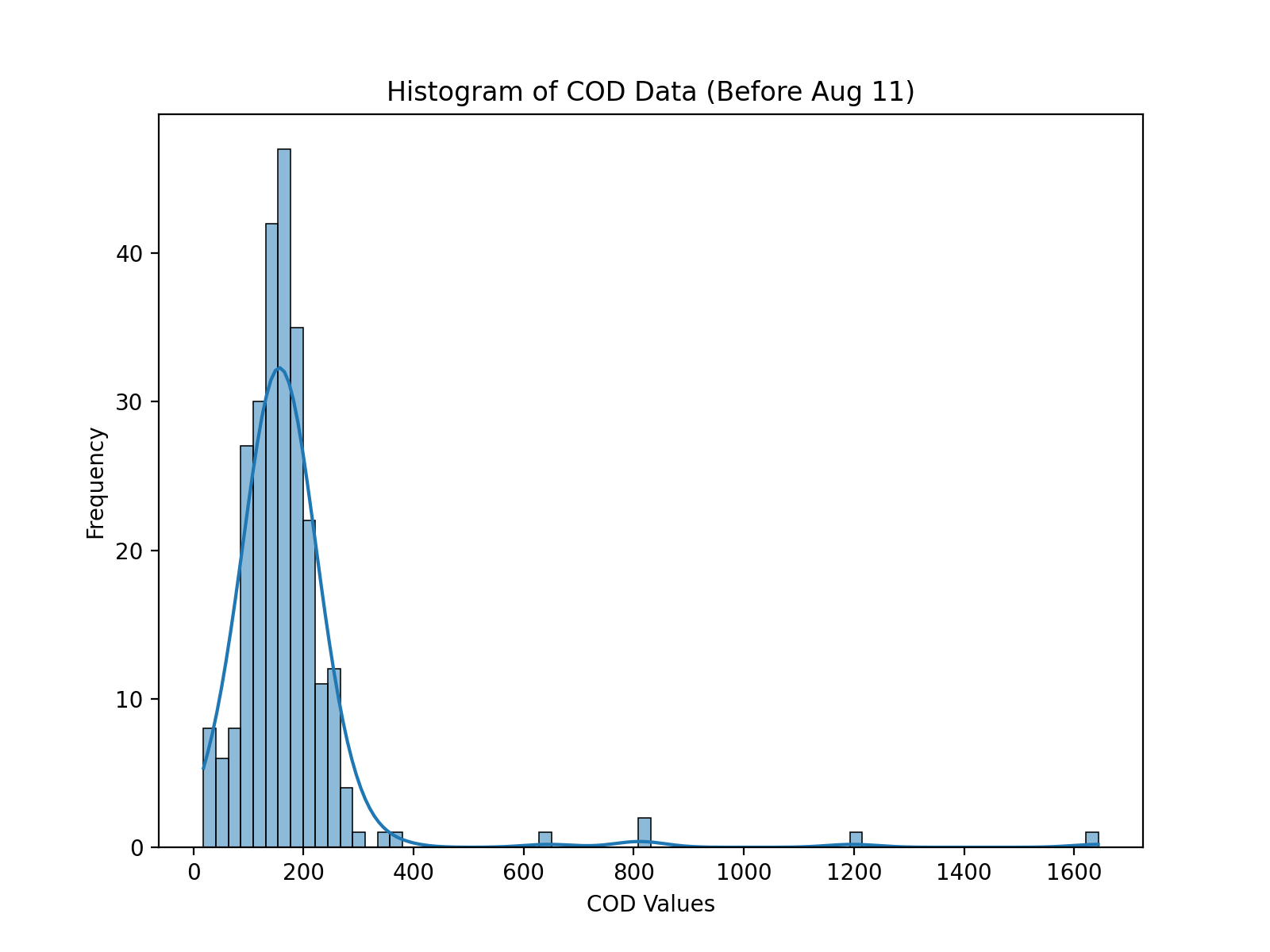


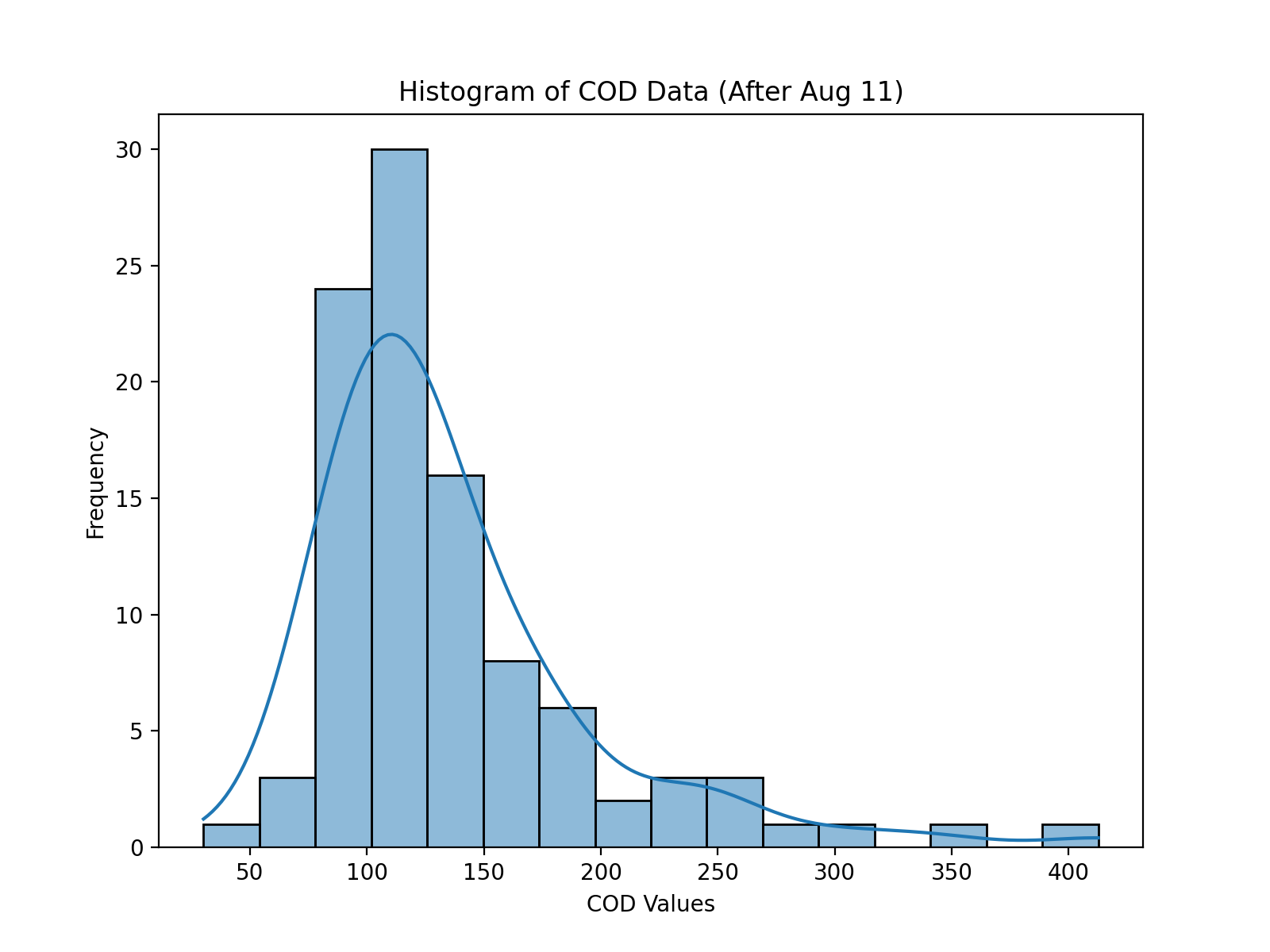
红色数据为8月11日前的数据分布，可以看出一天的分布较为平均，蓝色为11日后的分布，可以看出分布在24小时的尺度上呈现出明显的高低差，尤其表现在12与14时。

归一化之后的概率密度分布



归一化后的查看数据是否符合正态分布：





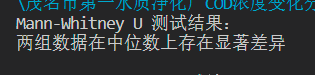
前后数据均不符合正态分布，

Mann-Whitney U 测试（Wilcoxon秩和检验）：

用途：比较两组独立样本的中位数是否相同。

原假设：两组数据的分布相同。

alpha = 0.05 # 设置显著性水平



出现这个错误的原因是 wilcoxon 函数要求两组样本具有相同的长度，以便执行配对样本的 Wilcoxon 符号秩检验。在您的代码中，group1 和 group2 的长度不一致，因此引发了异常。

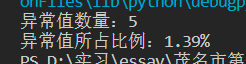
异常值处理：

标准差方法（Standard Deviation Method）：

计算数据的均值和标准差。

定义一个阈值，通常是均值加上或减去几个标准差的倍数。例如，2倍标准差通常用于识别异常值。

任何超出这个阈值的数据点都被视为异常值。

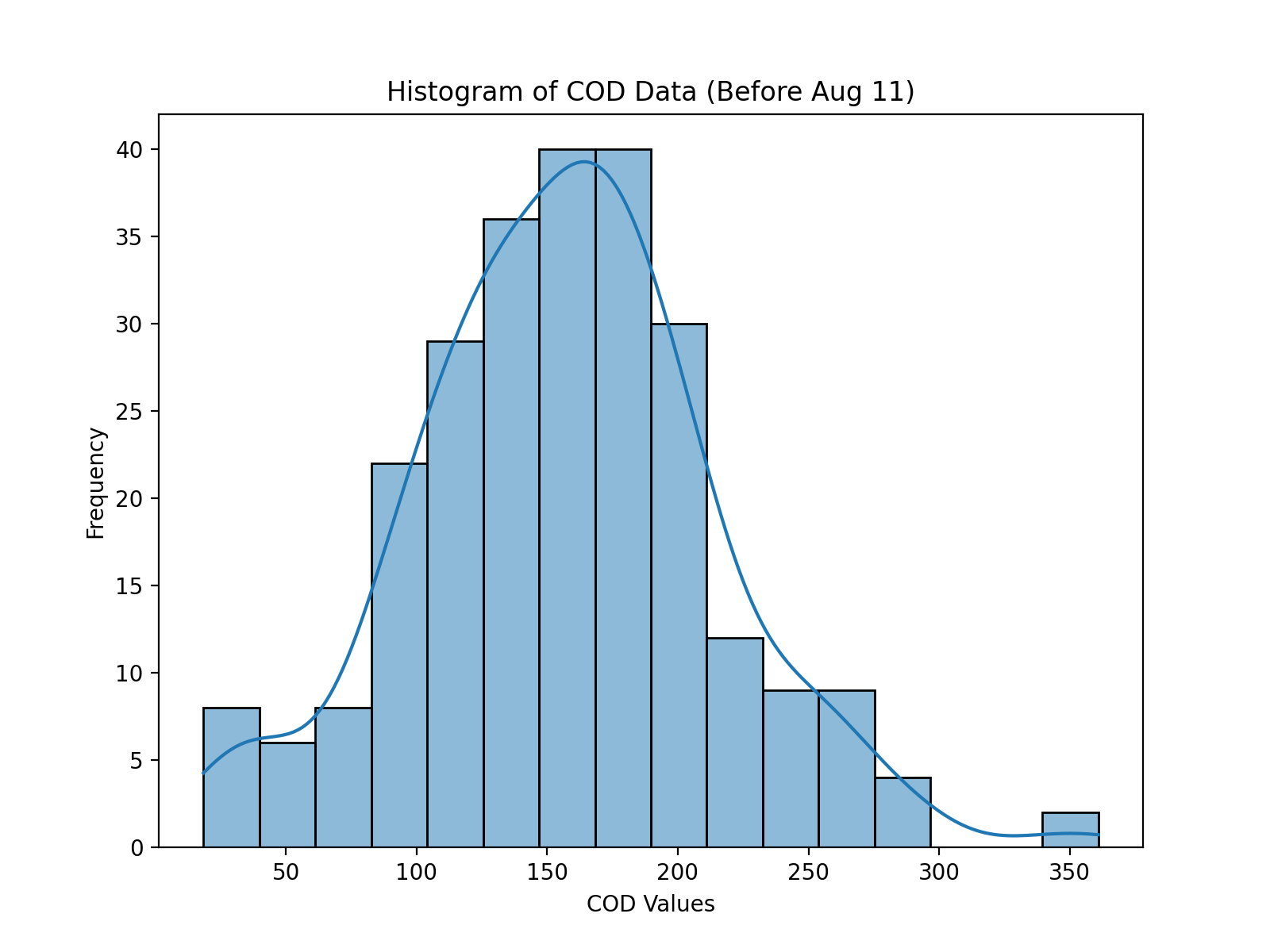


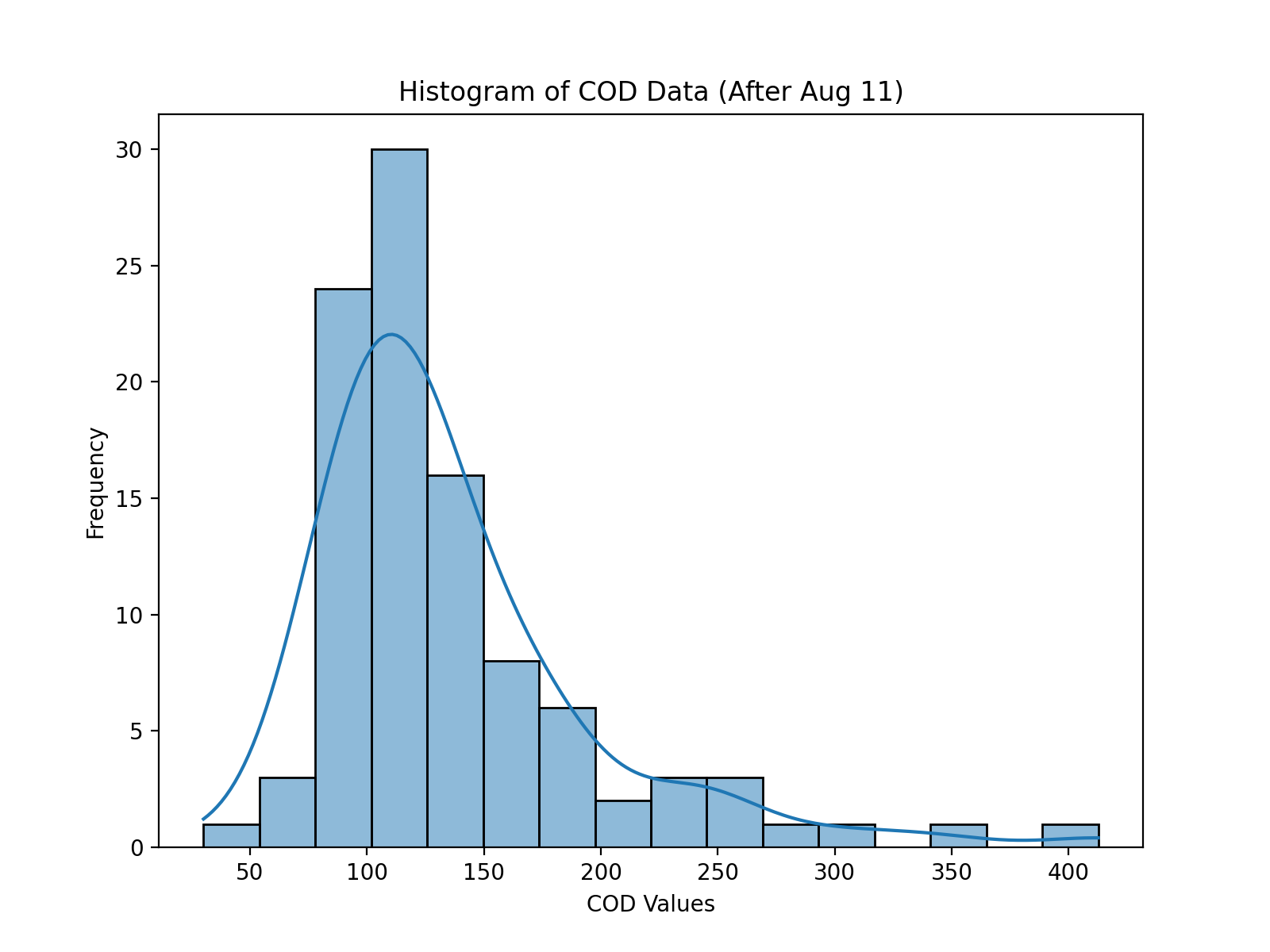
异常值处理后的正态分布：

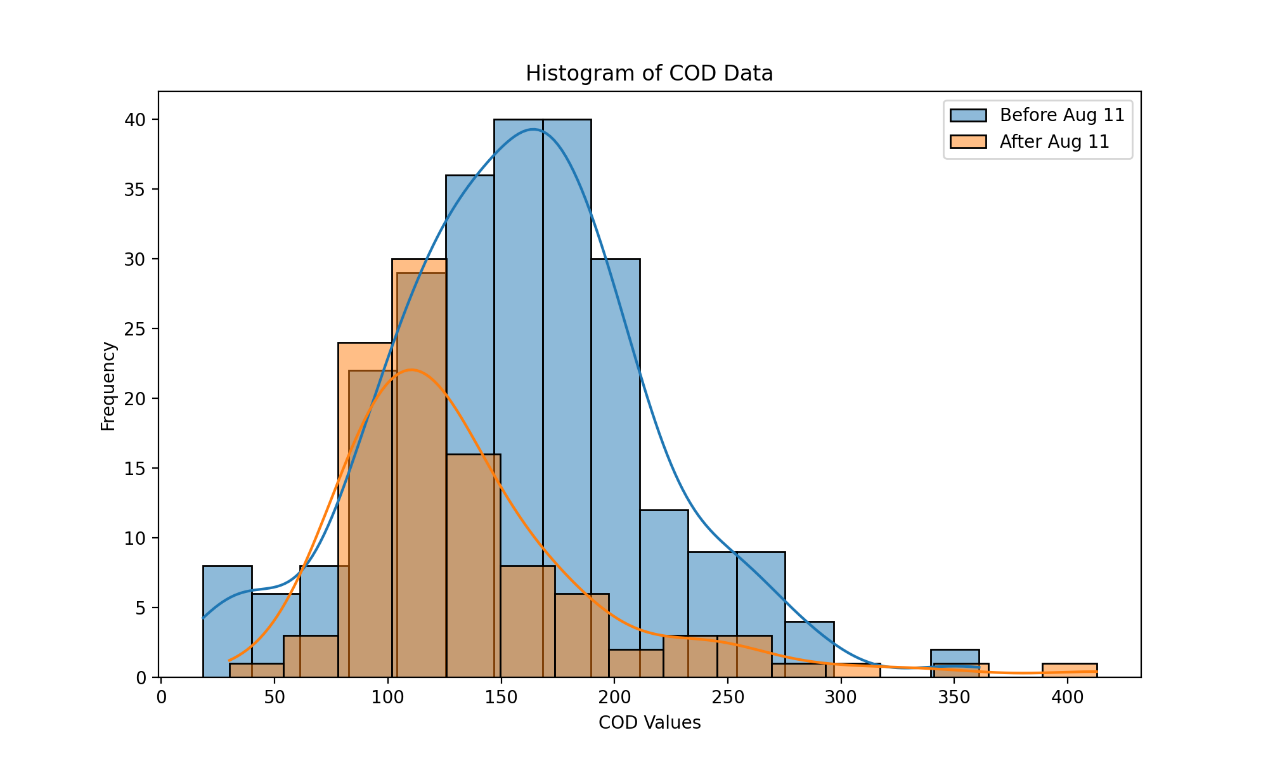
正态性检验 - Shapiro-Wilk 测试

alpha = 0.05 # 设置显著性水平

8月11日之前的数据符合正态分布







异常值处理后

