数据挖掘作业1

姓名：杨 亮

学号：2120151053

一、数据摘要

1、对标称属性，给出每个可能取值的频数

标称变量“季节”的频数：

{'autumn': 40, 'spring': 53, 'winter': 62, 'summer': 45}

标称变量“河流大小”的频数：

{'small': 71, 'large': 45, 'medium': 84}

标称变量“河流速度”的频数：

{'high': 84, 'medium': 83, 'low': 33}

2、数值属性，最大、最小、均值、中位数、四分位数及缺失值的个数

'max'表示最大值

'min'表示最小值

'mean'表示最均值

'mid'表示中位数

'quantile'表示四分位数

'max'表示最大值

'lost\_num'表示缺失值的个数

数值属性'mxPH'：

('mxPH', {'lost\_num': 1, 'min': 5.6, 'max': 9.7, 'quantile': {'Q1': 7.7, 'Q3': 8.4, 'Q2': 8.06}, 'mid': 8.06, 'mean': 8.0117336683417086})

数值属性'mnO2'：

('mnO2', {'lost\_num': 2, 'min': 1.5, 'max': 13.4, 'quantile': {'Q1': 7.7, 'Q3': 10.8, 'Q2': 9.8}, 'mid': 9.8, 'mean': 9.1177777777777713})

数值属性'CI'：

('CI', {'lost\_num': 10, 'min': 0.222, 'max': 391.5, 'quantile': {'Q1': 10.96525, 'Q3': 57.96875, 'Q2': 32.730000000000004}, 'mid': 32.730000000000004, 'mean': 43.636278842105256})

数值属性'NO3'：

('NO3', {'lost\_num': 2, 'min': 0.05, 'max': 45.65, 'quantile': {'Q1': 1.28775, 'Q3': 4.4712499999999995, 'Q2': 2.675}, 'mid': 2.675, 'mean': 3.2823888888888906})

数值属性'NH4'：

('NH4', {'lost\_num': 2, 'min': 5.0, 'max': 24064.0, 'quantile': {'Q1': 38.02775, 'Q3': 227.8875075, 'Q2': 103.1665}, 'mid': 103.1665, 'mean': 501.2958283838384})

数值属性'oPO4'：

('oPO4', {'lost\_num': 2, 'min': 1.0, 'max': 564.59998, 'quantile': {'Q1': 15.36675, 'Q3': 100.5, 'Q2': 40.15}, 'mid': 40.15, 'mean': 73.590595959595973})

数值属性'PO4'：

('PO4', {'lost\_num': 2, 'min': 1.0, 'max': 771.59998, 'quantile': {'Q1': 40.50025, 'Q3': 214.375, 'Q2': 103.2855}, 'mid': 103.2855, 'mean': 137.88210095959593})

数值属性'Chla'：

('Chla', {'lost\_num': 12, 'min': 0.2, 'max': 110.456, 'quantile': {'Q1': 2.0, 'Q3': 18.322499999999998, 'Q2': 5.475}, 'mid': 5.475, 'mean': 13.971196808510641})

数值属性'a1'：

('a1', {'lost\_num': 0, 'min': 0.0, 'max': 89.8, 'quantile': {'Q1': 1.5, 'Q3': 24.8, 'Q2': 6.95}, 'mid': 6.95, 'mean': 16.923500000000001})

数值属性'a2：

('a2', {'lost\_num': 0, 'min': 0.0, 'max': 72.6, 'quantile': {'Q1': 0.0, 'Q3': 11.525, 'Q2': 3.0}, 'mid': 3.0, 'mean': 7.4585000000000017})

数值属性'a3：

('a3', {'lost\_num': 0, 'min': 0.0, 'max': 42.8, 'quantile': {'Q1': 0.0, 'Q3': 4.975, 'Q2': 1.55}, 'mid': 1.55, 'mean': 4.3095000000000008})

数值属性'a4：

('a4', {'lost\_num': 0, 'min': 0.0, 'max': 44.6, 'quantile': {'Q1': 0.0, 'Q3': 2.4, 'Q2': 0.0}, 'mid': 0.0, 'mean': 1.9925000000000004})

数值属性'a5：

('a5', {'lost\_num': 0, 'min': 0.0, 'max': 44.4, 'quantile': {'Q1': 0.0, 'Q3': 7.5, 'Q2': 1.9}, 'mid': 1.9, 'mean': 5.0644999999999989})

数值属性'a6：

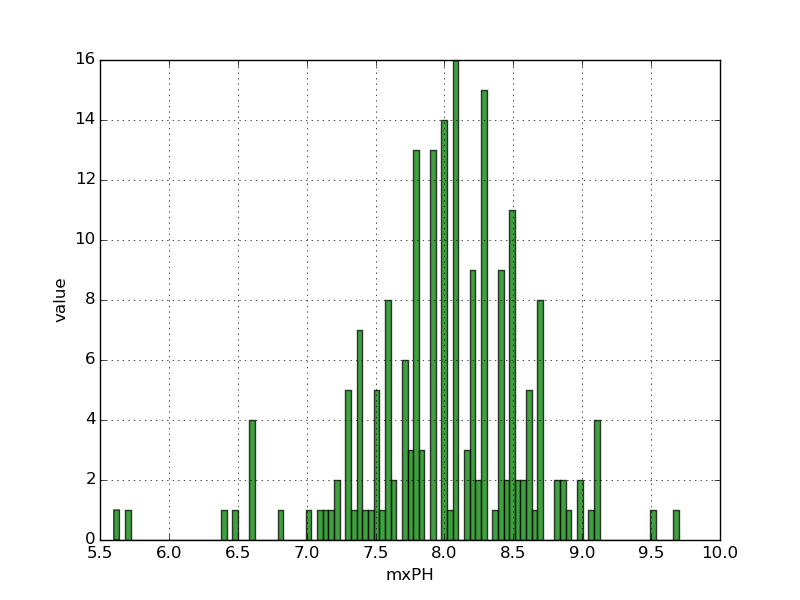
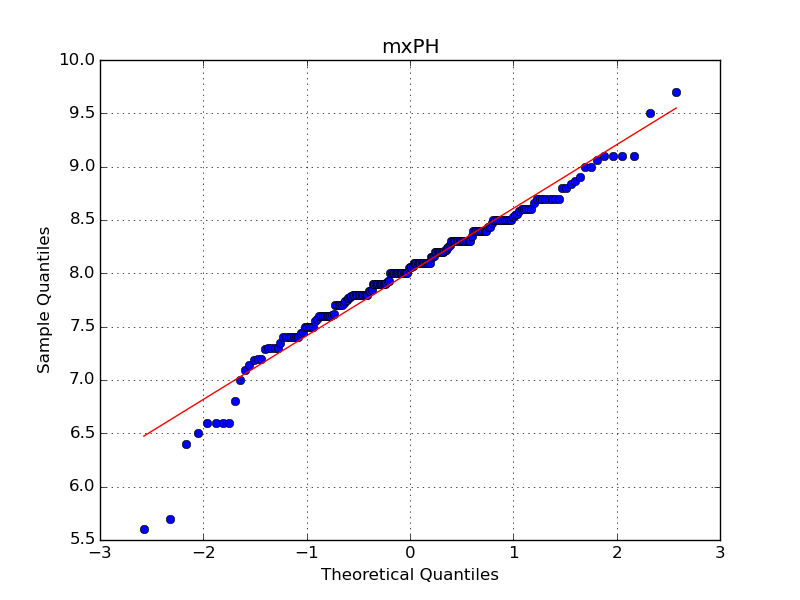
('a6', {'lost\_num': 0, 'min': 0.0, 'max': 77.6, 'quantile': {'Q1': 0.0, 'Q3': 6.975, 'Q2': 0.0}, 'mid': 0.0, 'mean': 5.9639999999999995})

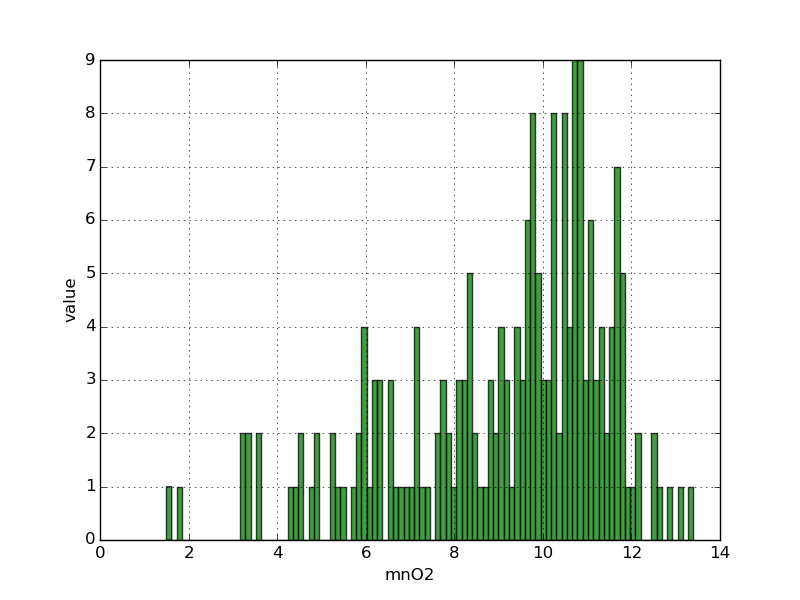
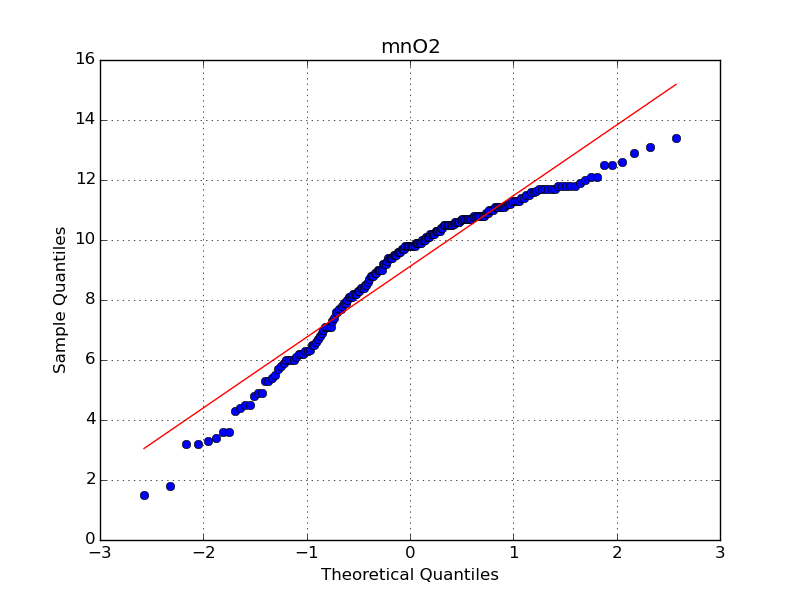
数值属性'a7：

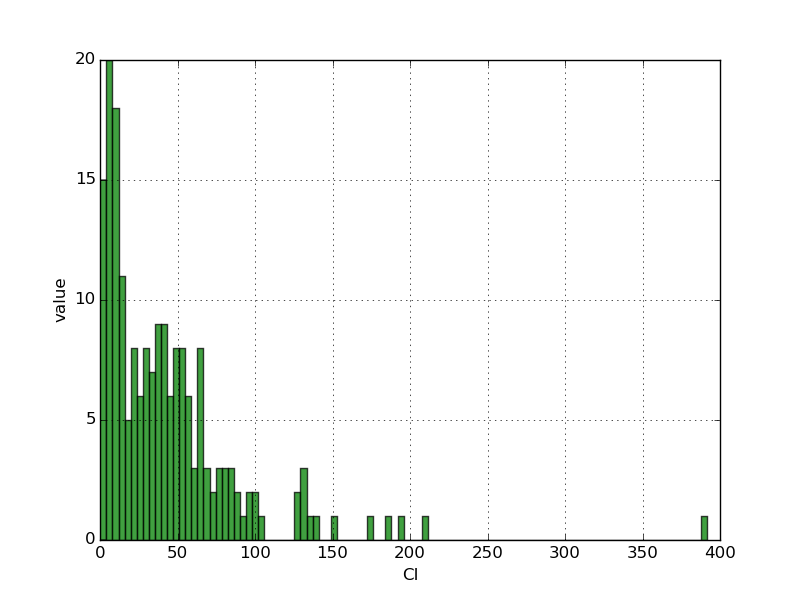
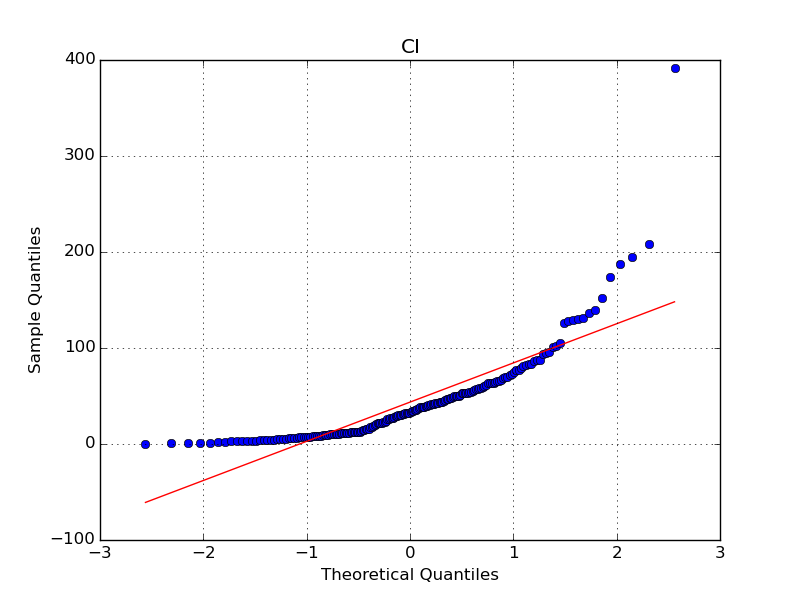
('a7', {'lost\_num': 0, 'min': 0.0, 'max': 31.6, 'quantile': {'Q1': 0.0, 'Q3': 2.4, 'Q2': 1.0}, 'mid': 1.0, 'mean': 2.4955000000000007})

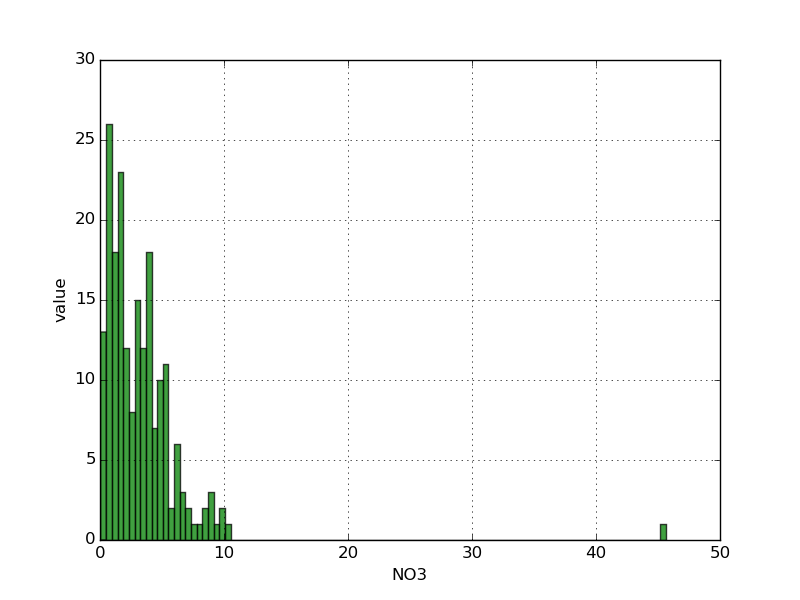
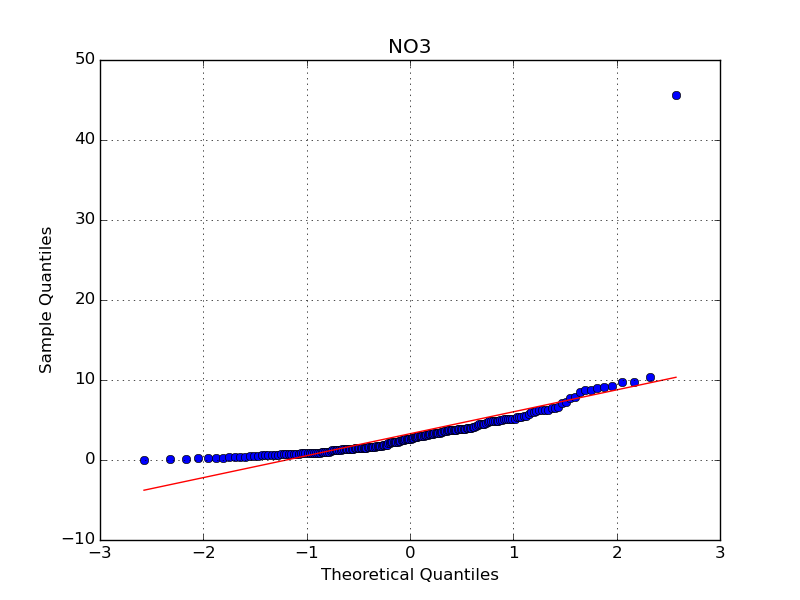
二、数据的可视化

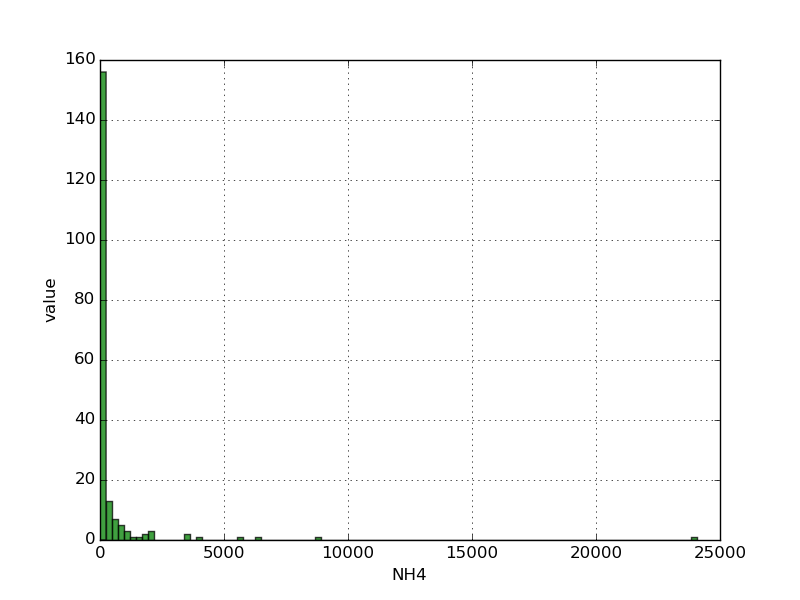
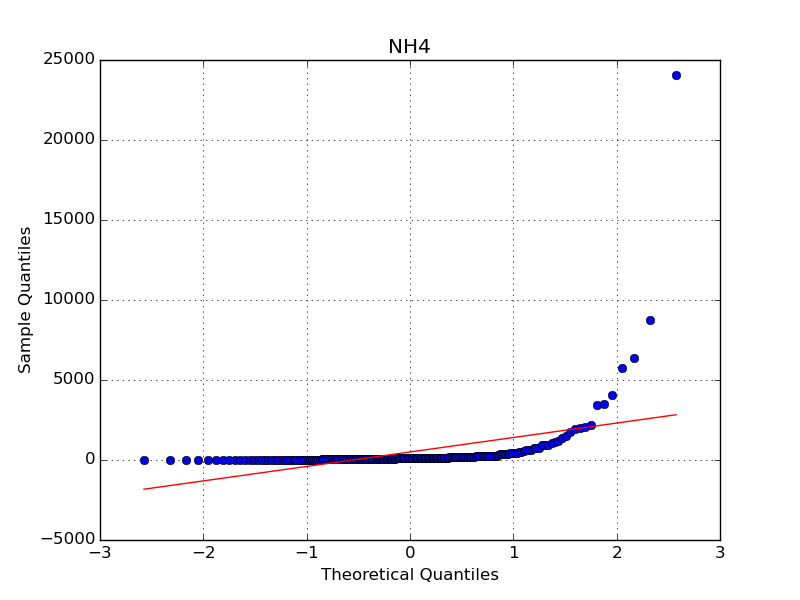
1、直方图及qq图检验

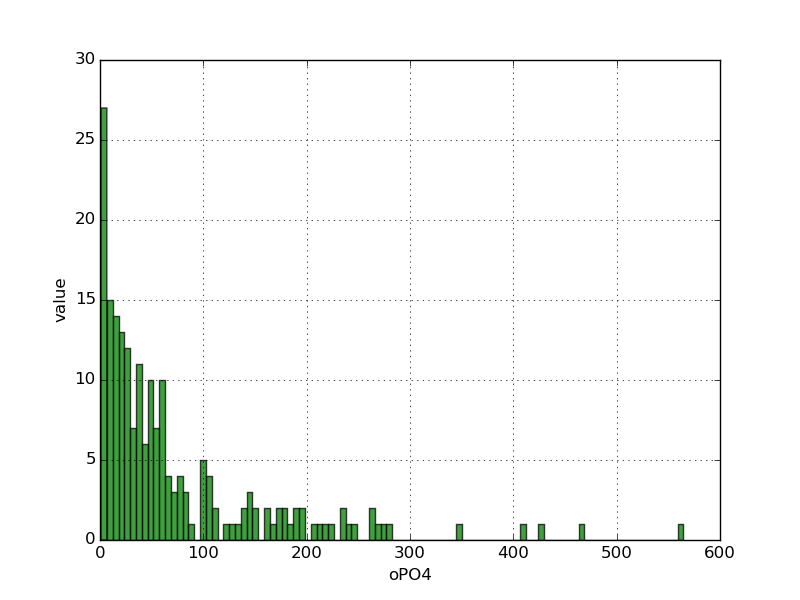
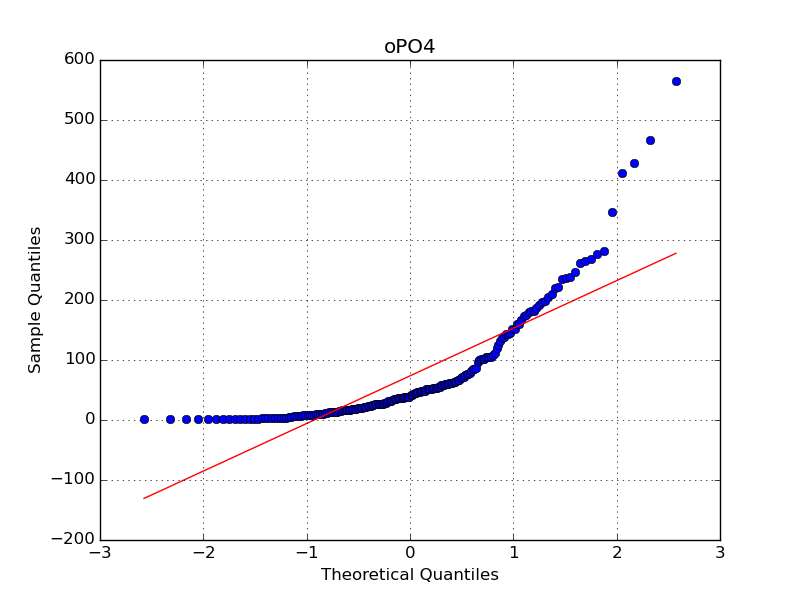
 

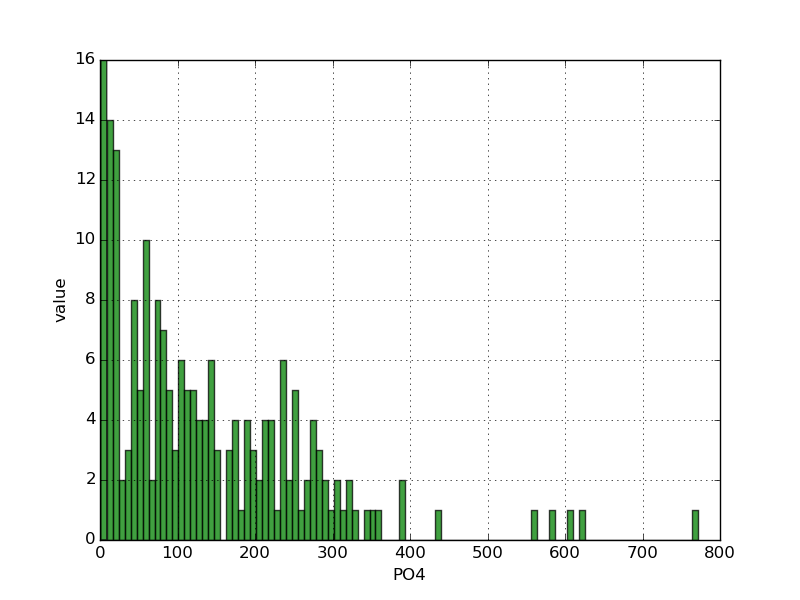
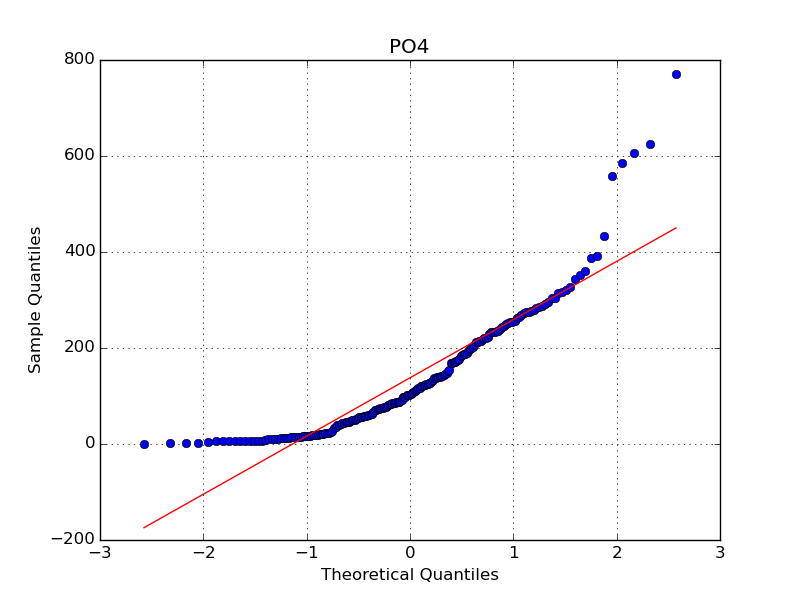
 

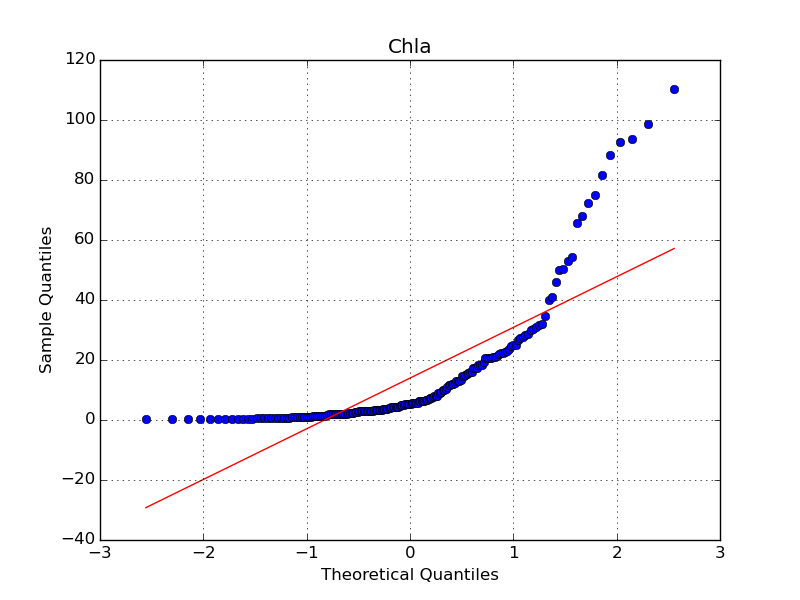
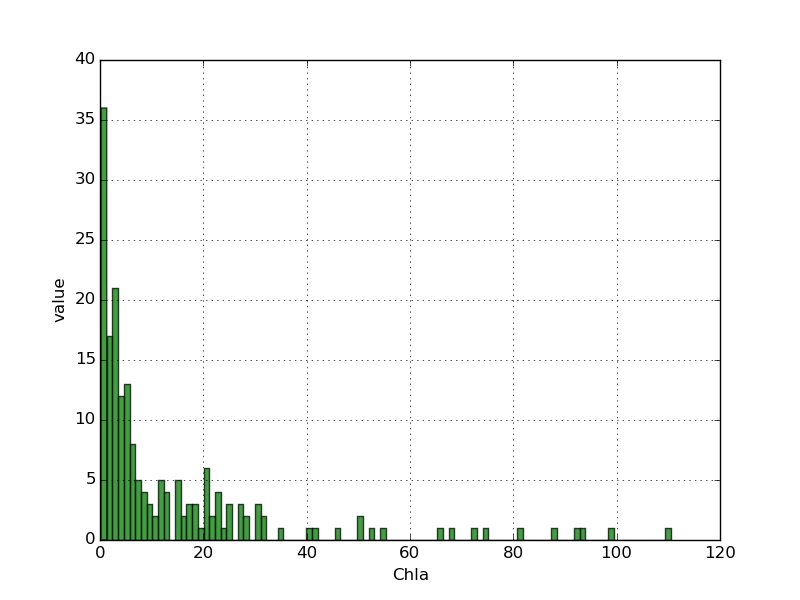
 

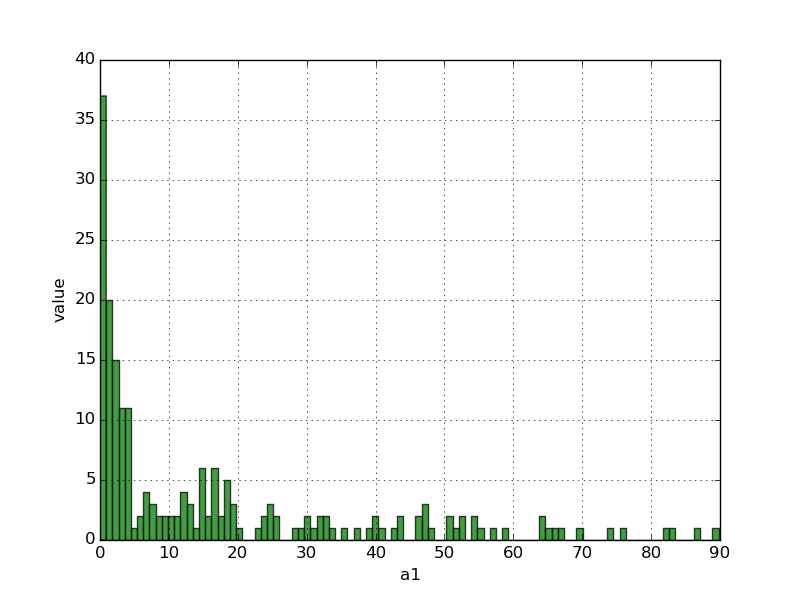
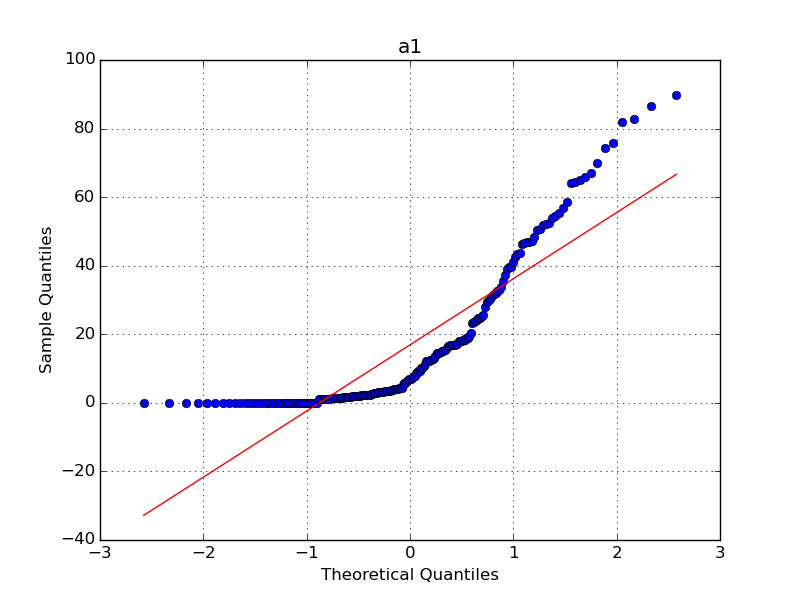
 

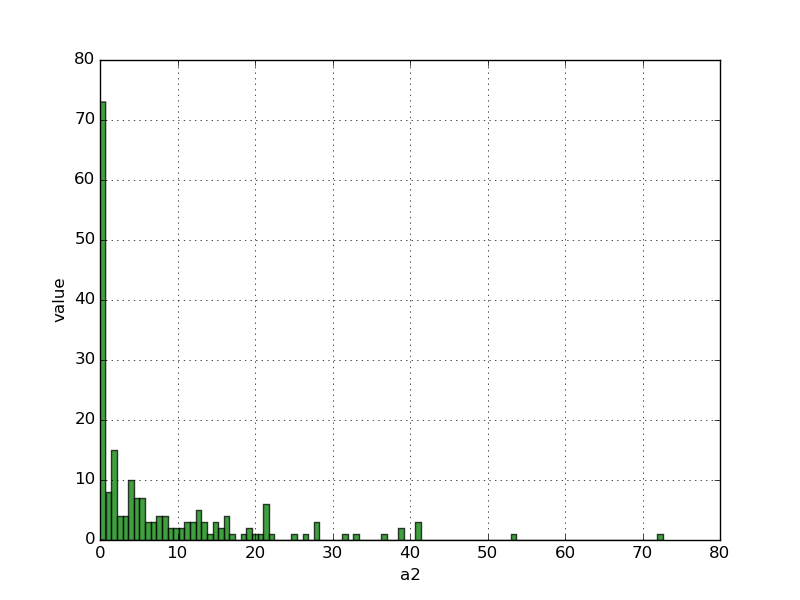
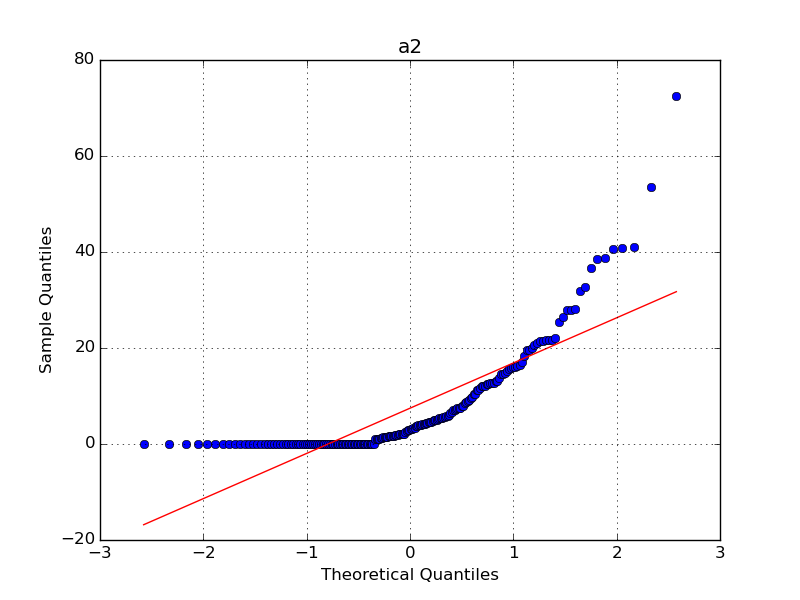
 

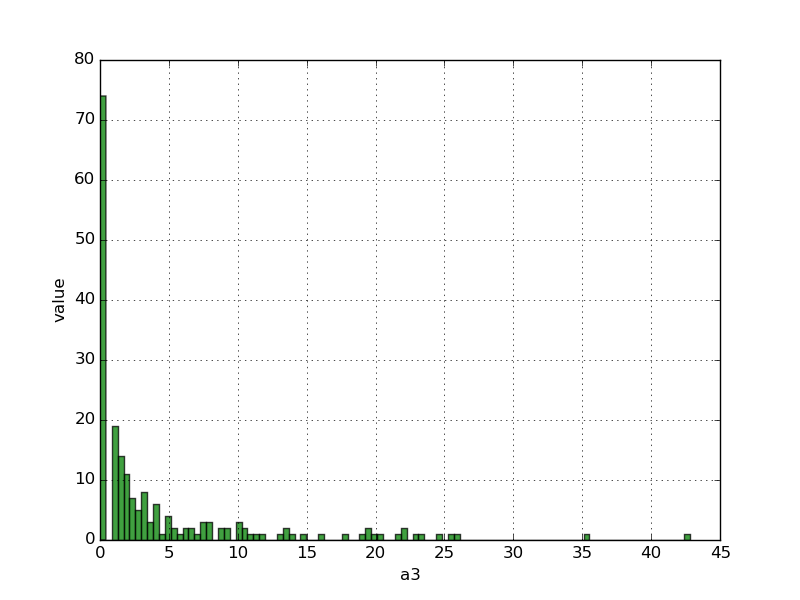
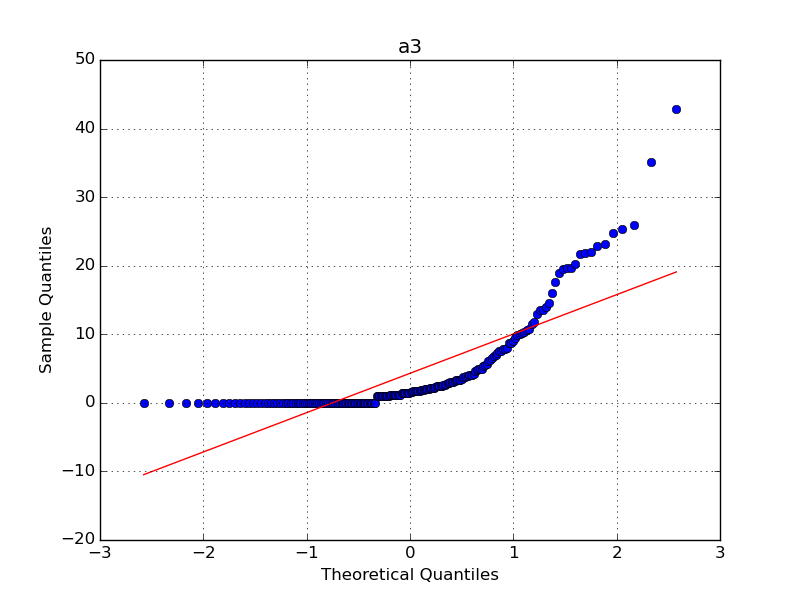
 

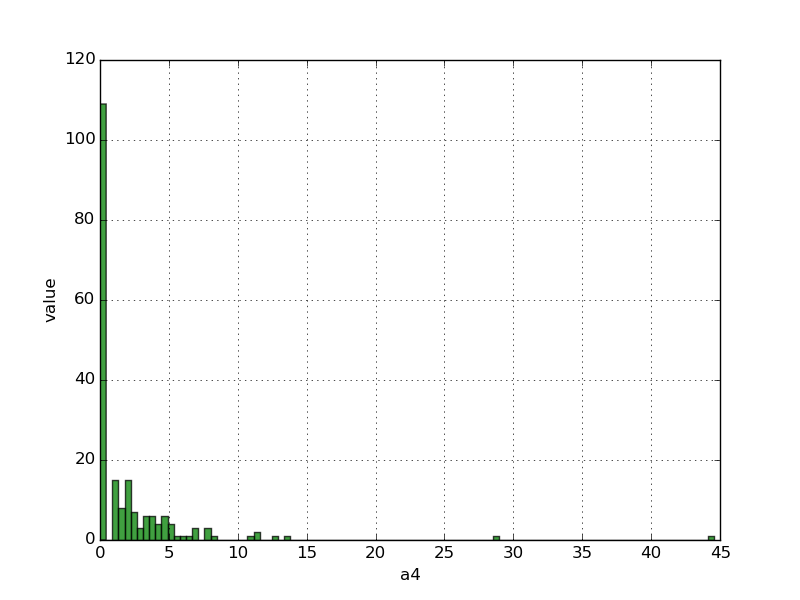
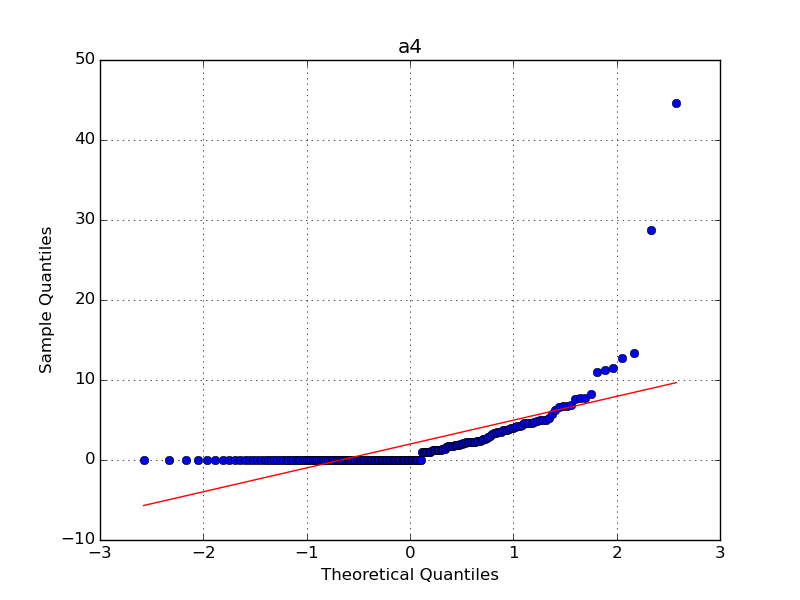
 

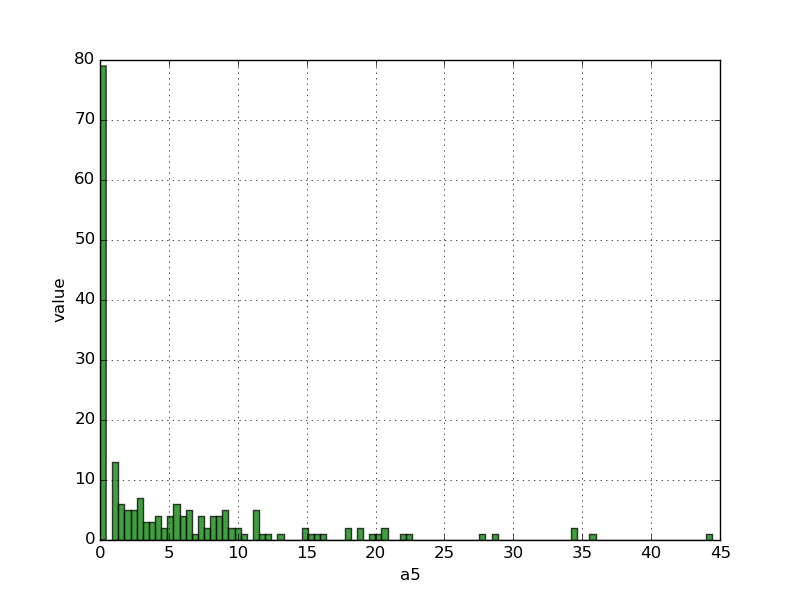
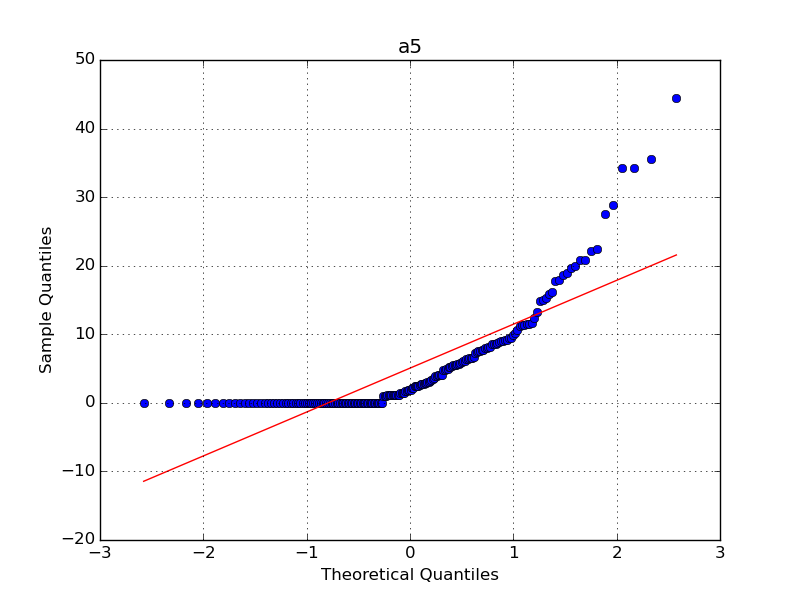


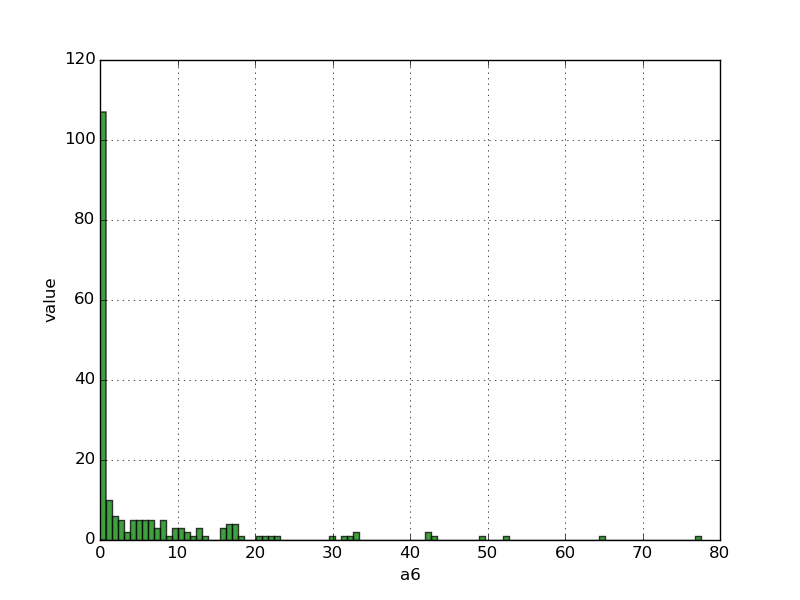
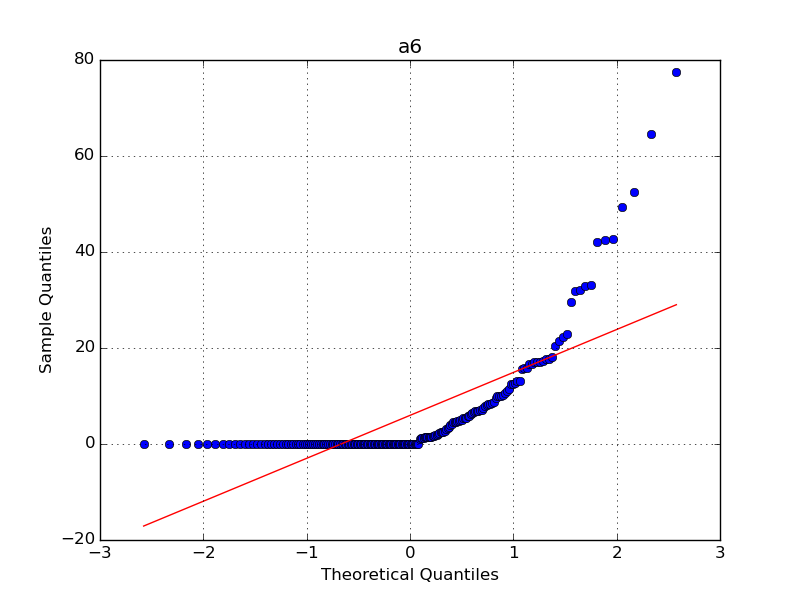
 

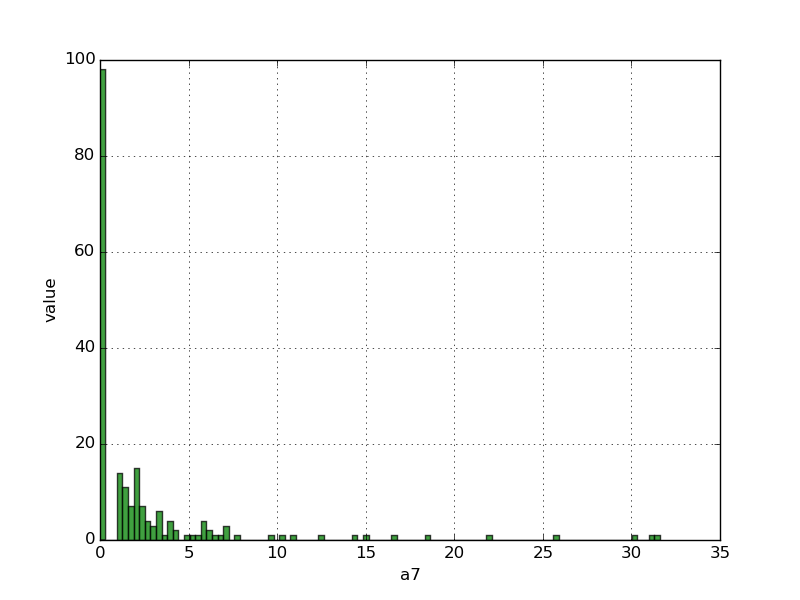
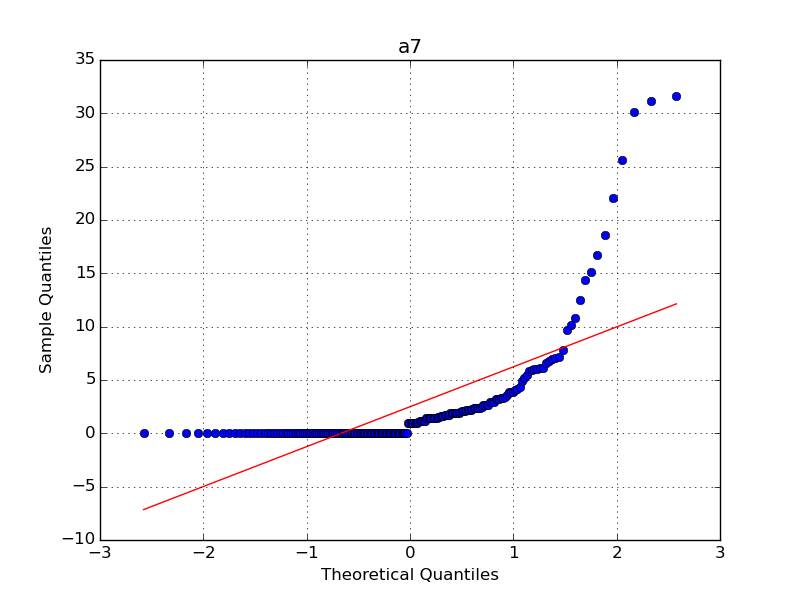
 

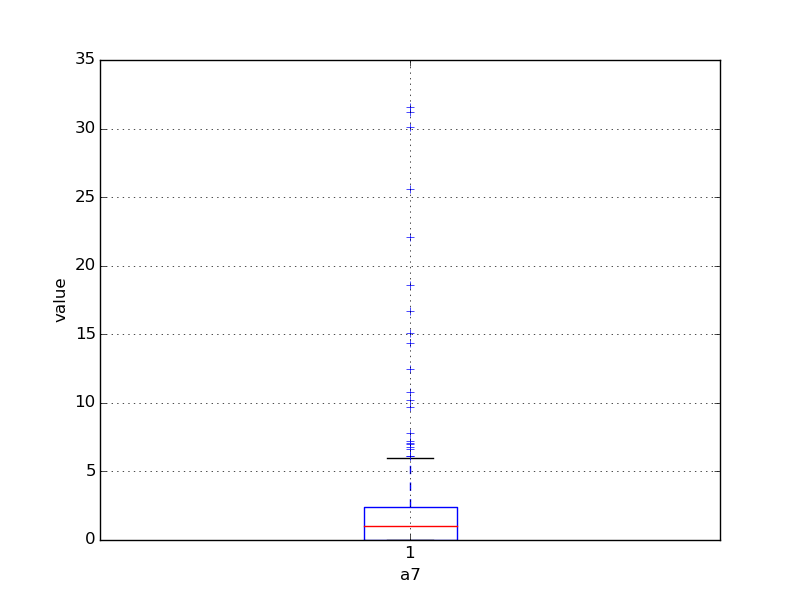
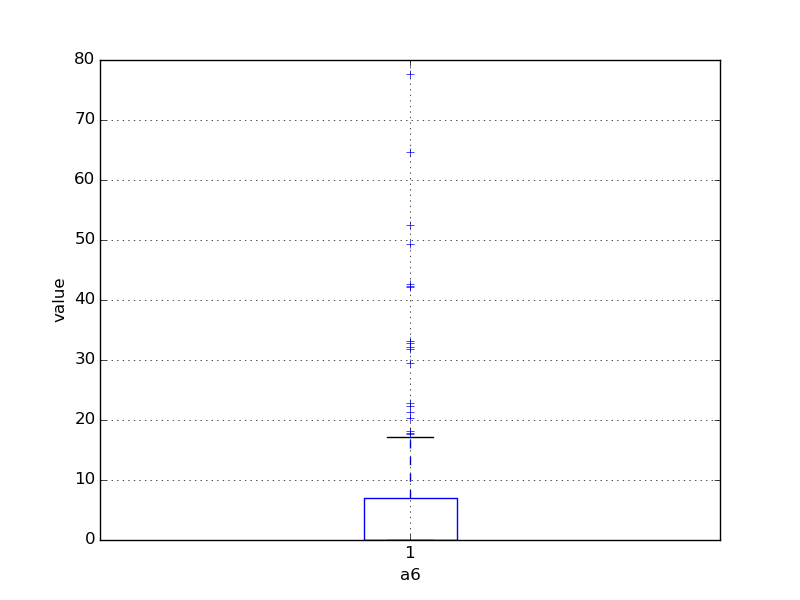
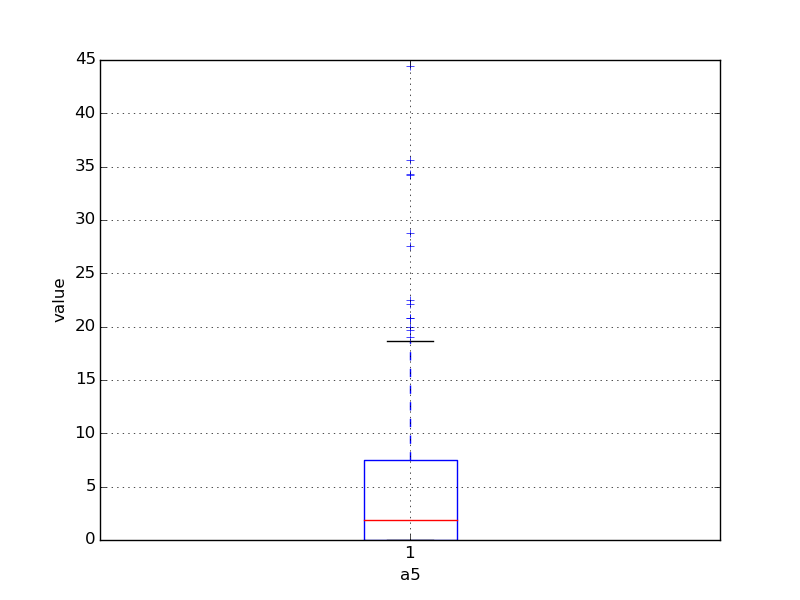
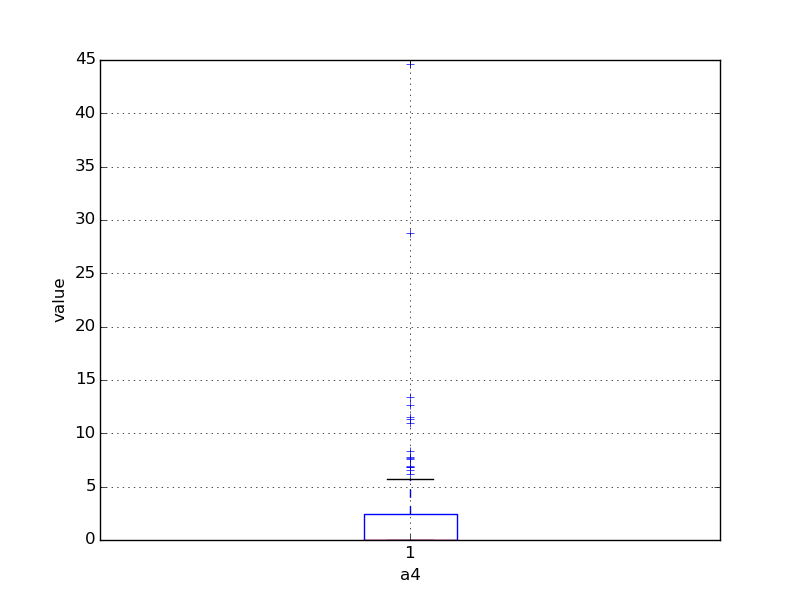
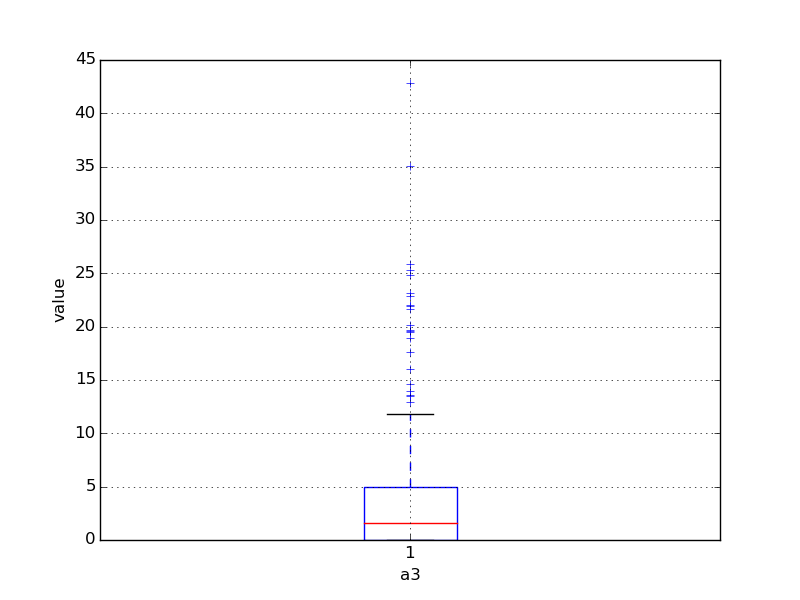
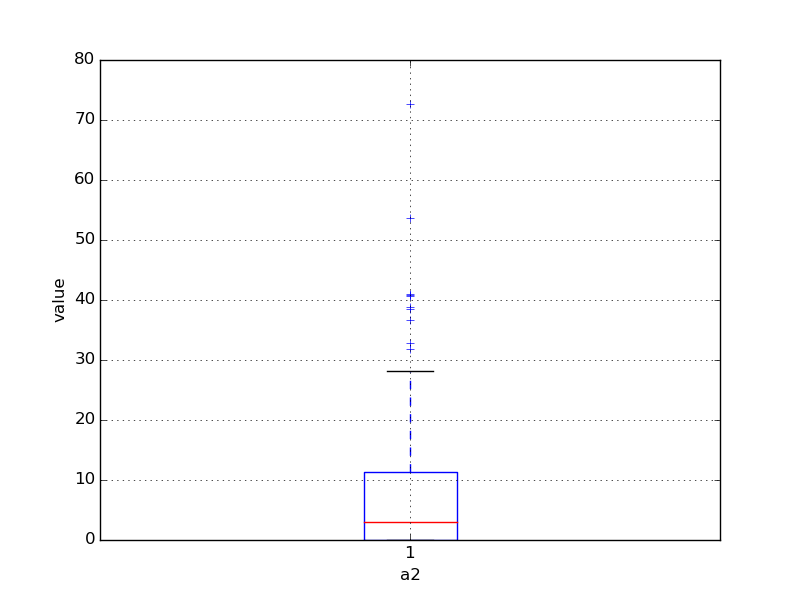
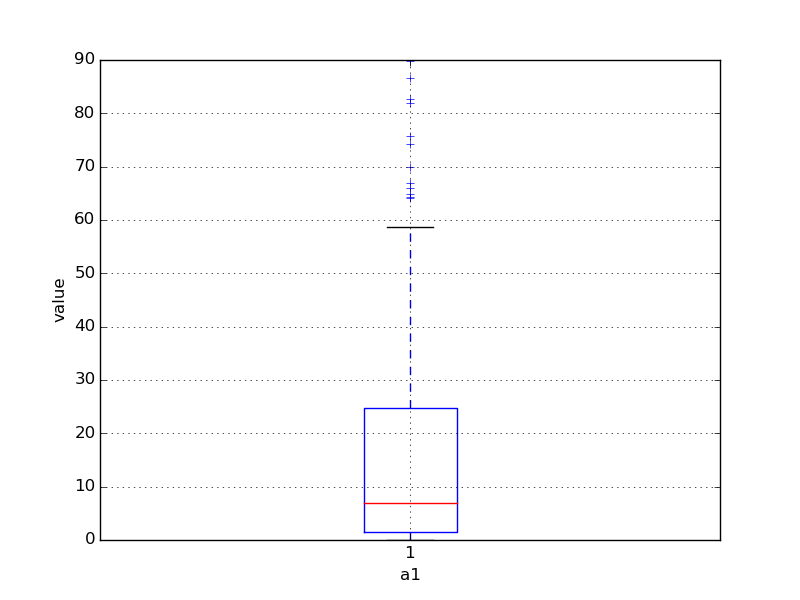
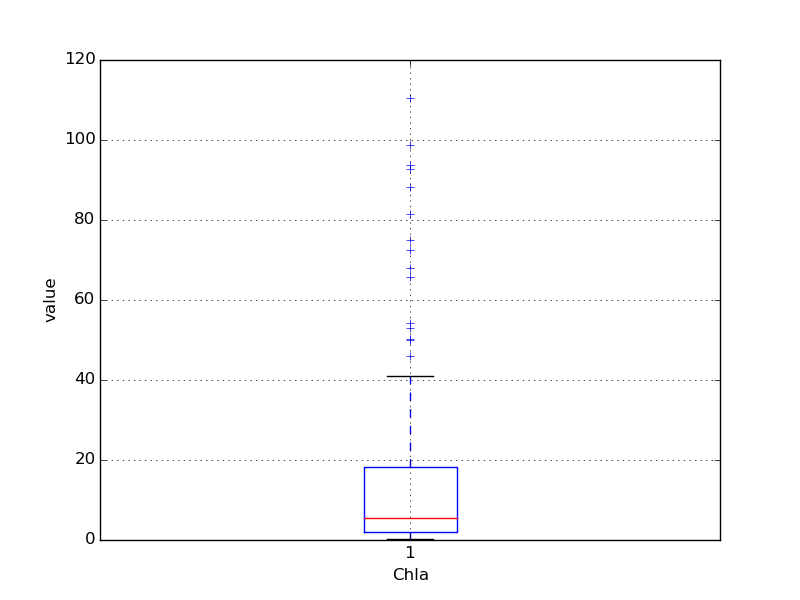
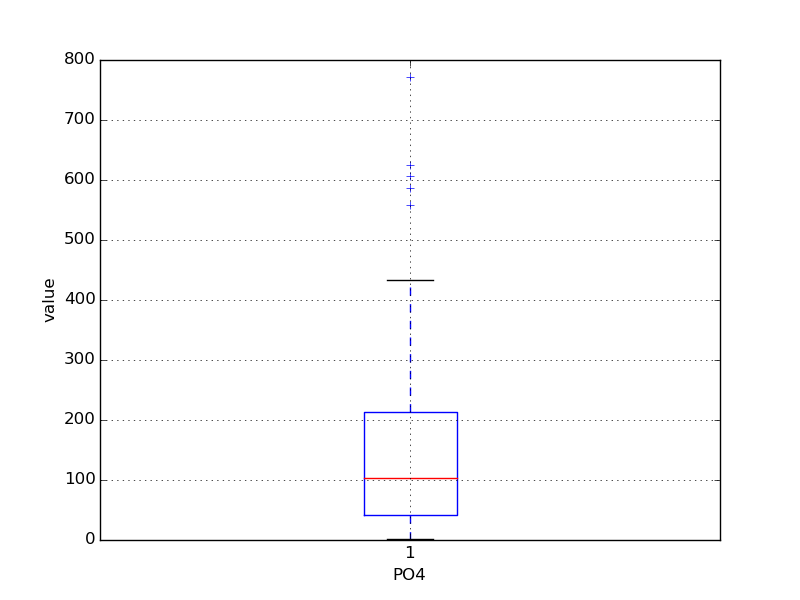
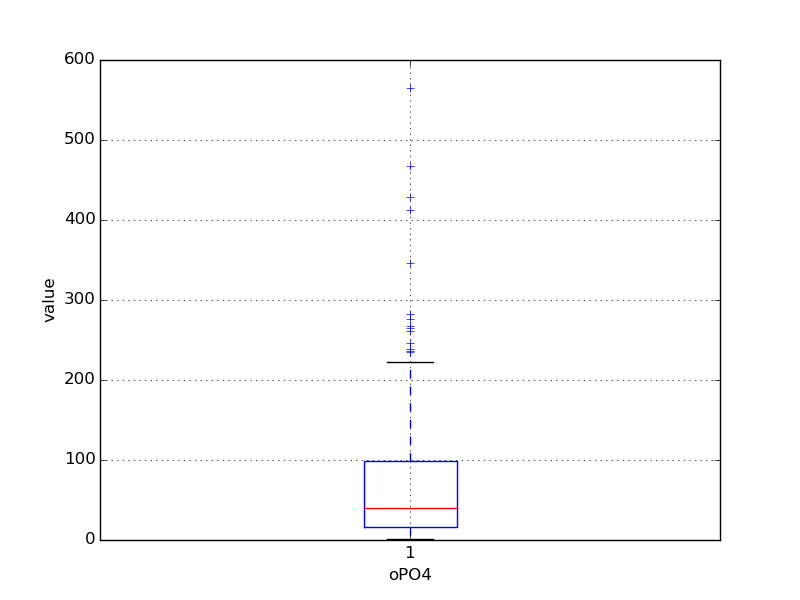
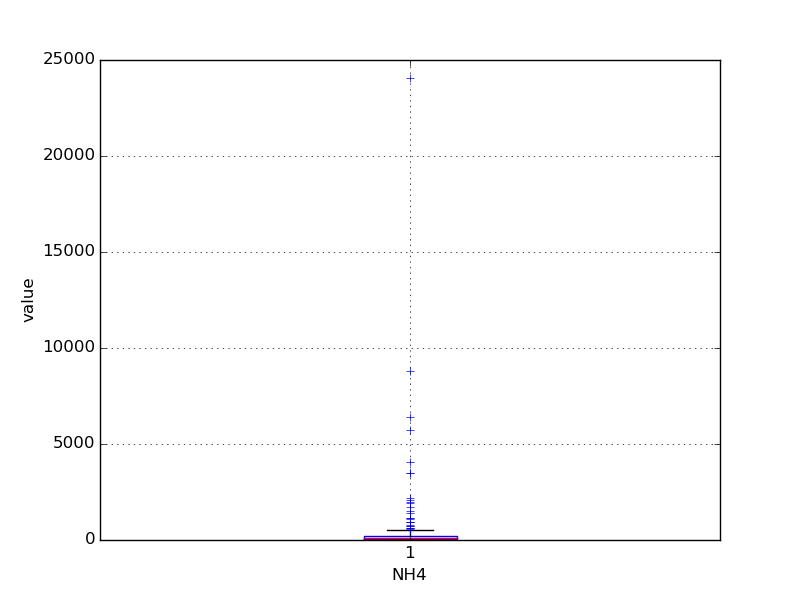
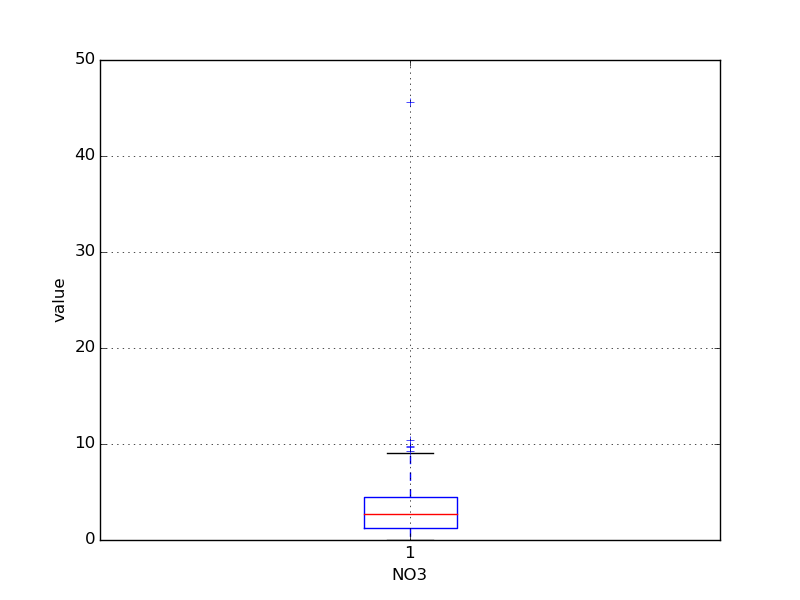
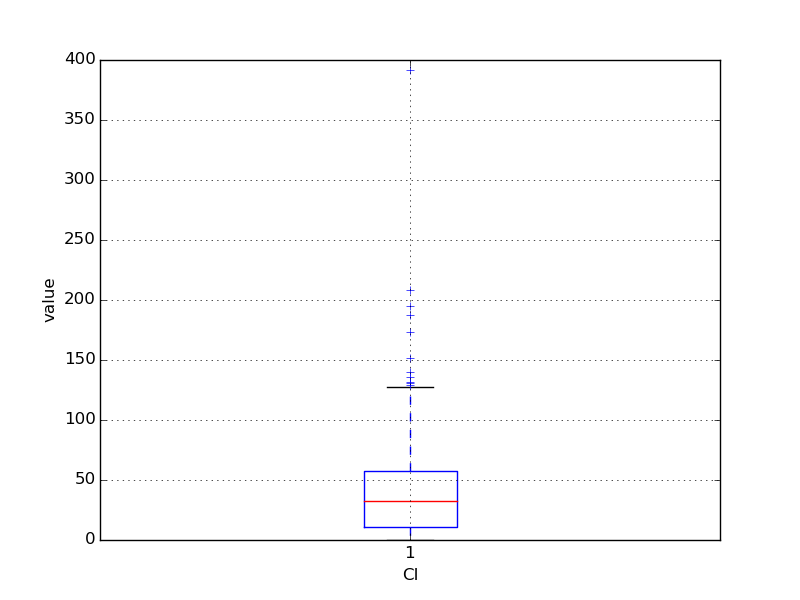
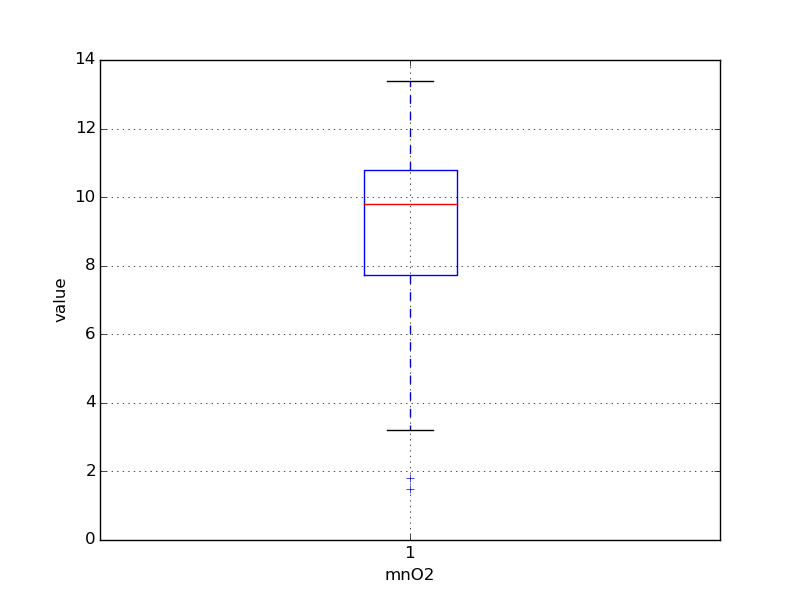
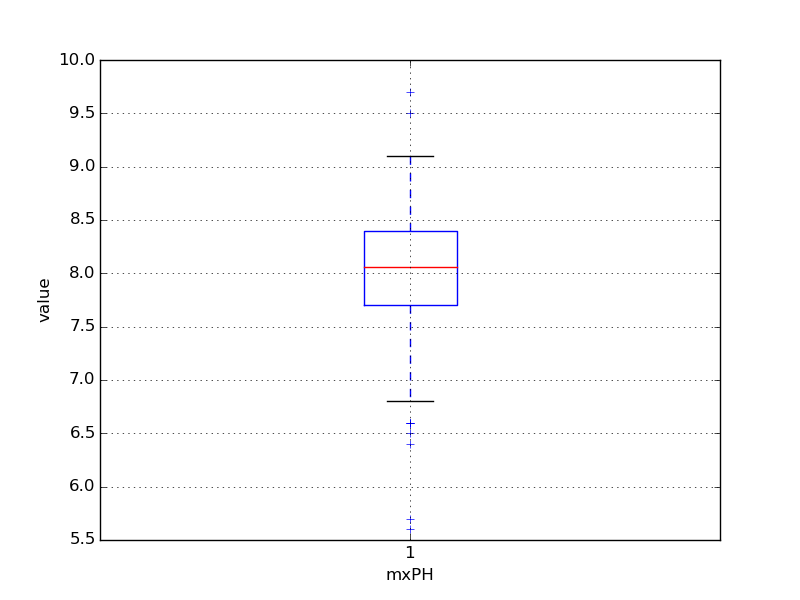
 

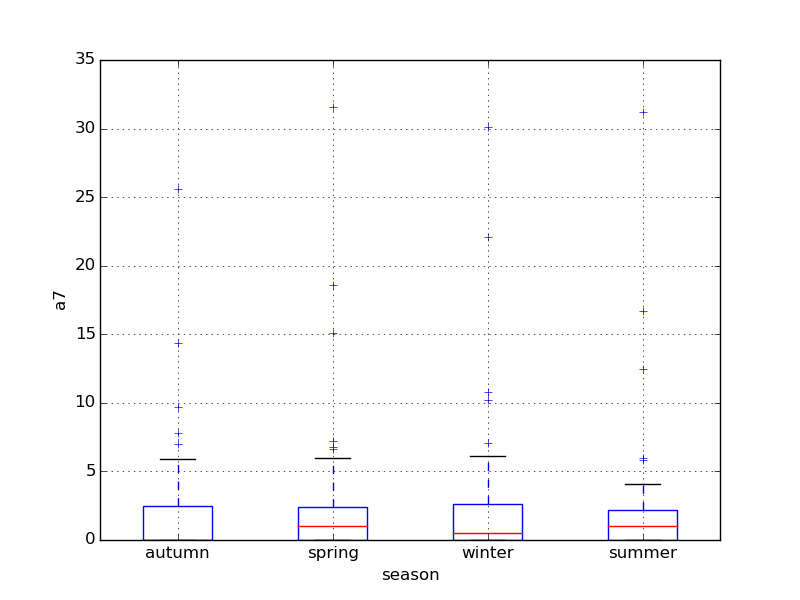
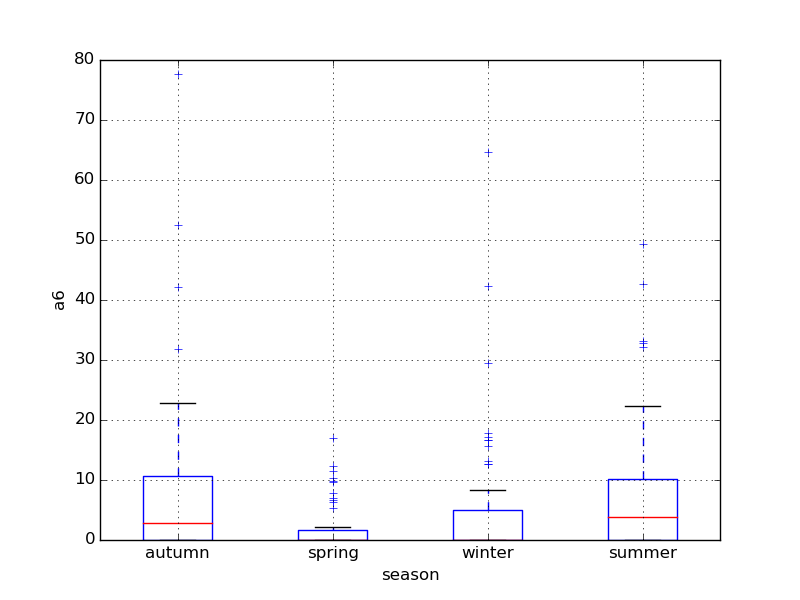
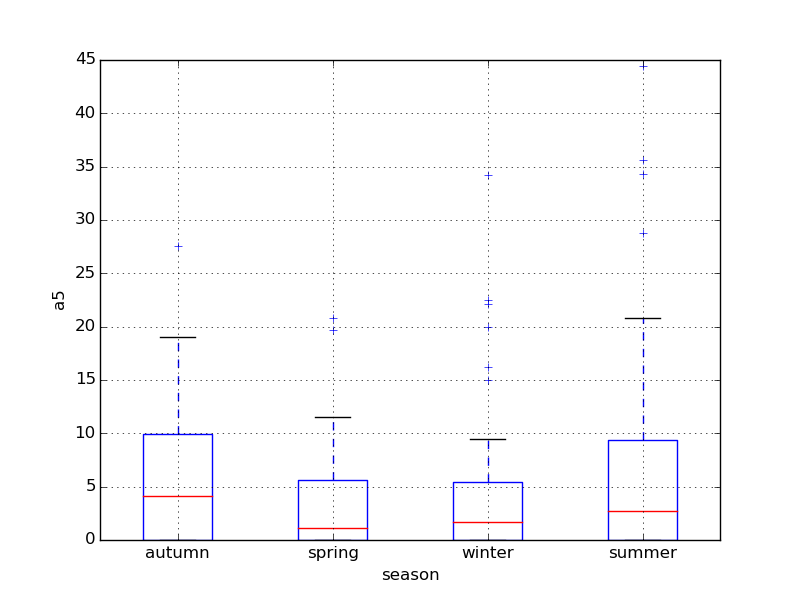
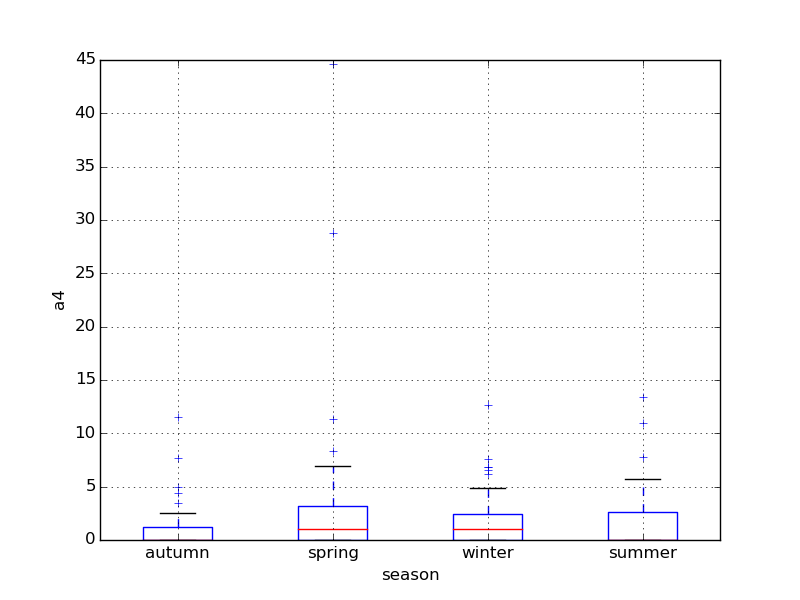
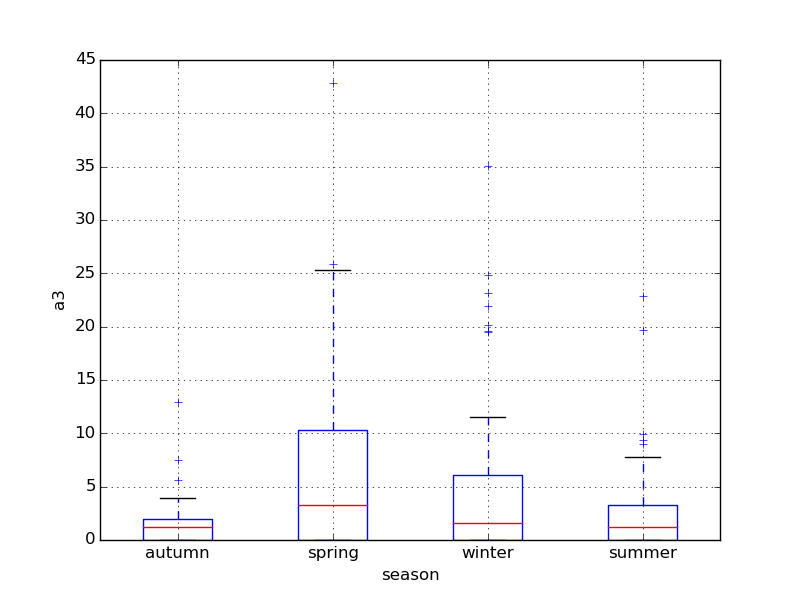
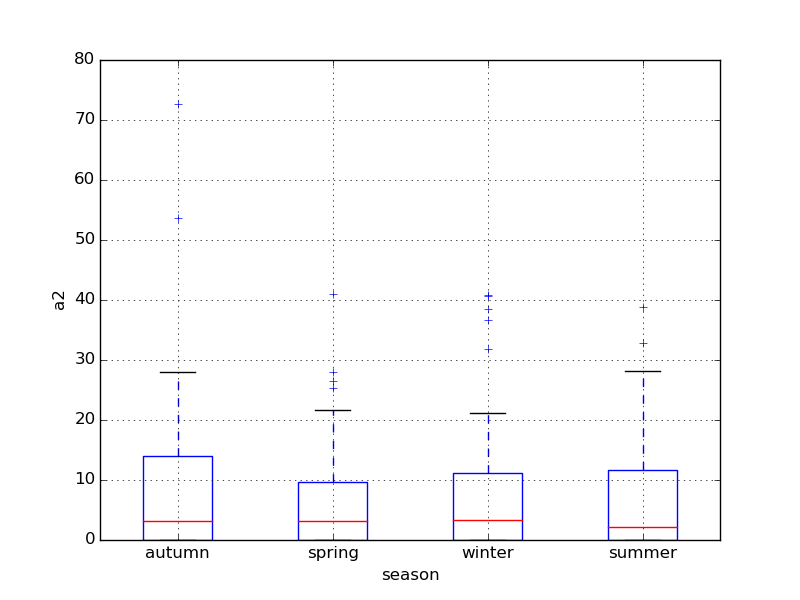
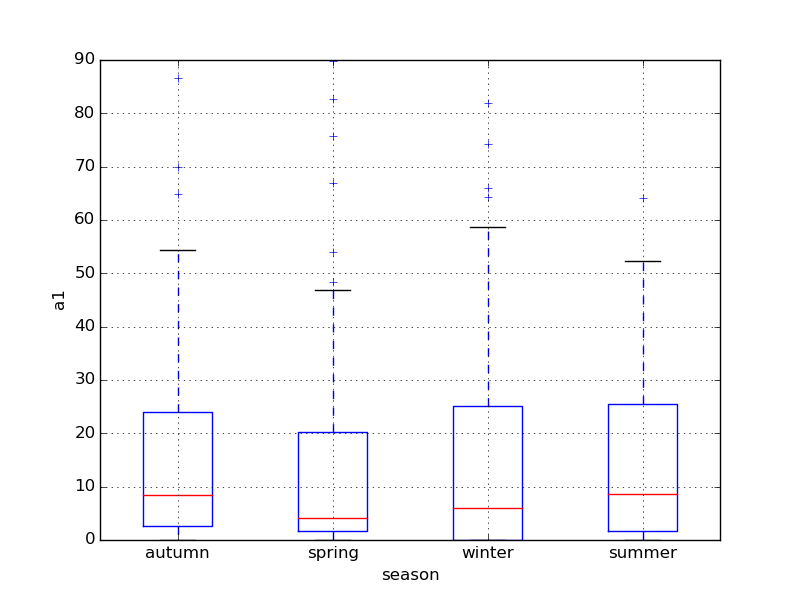
 

2、盒图

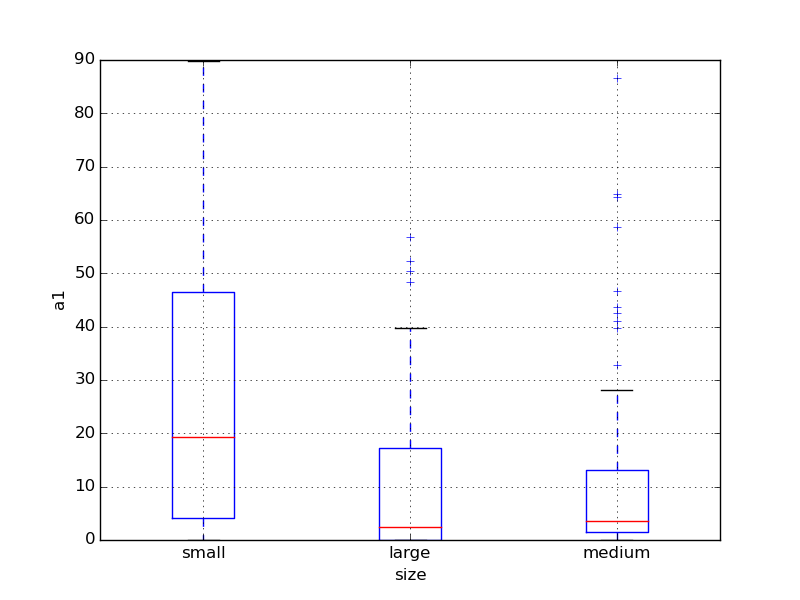
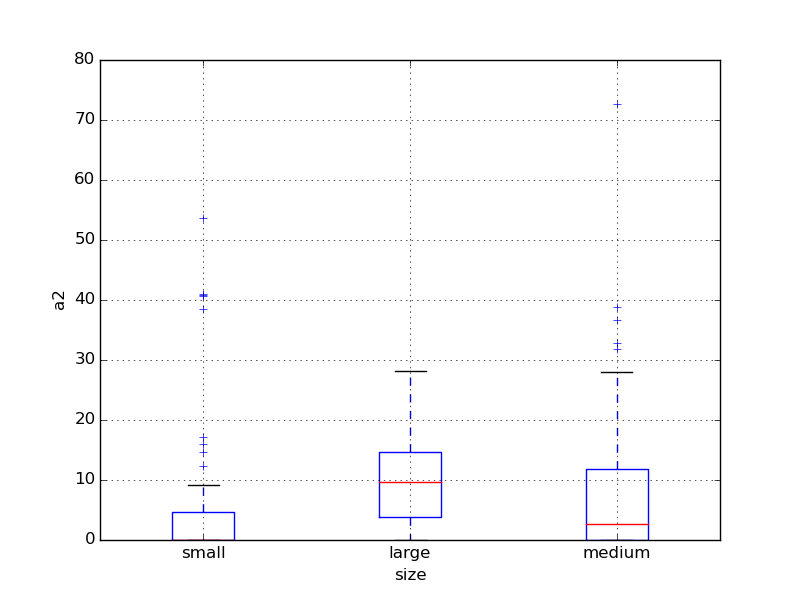
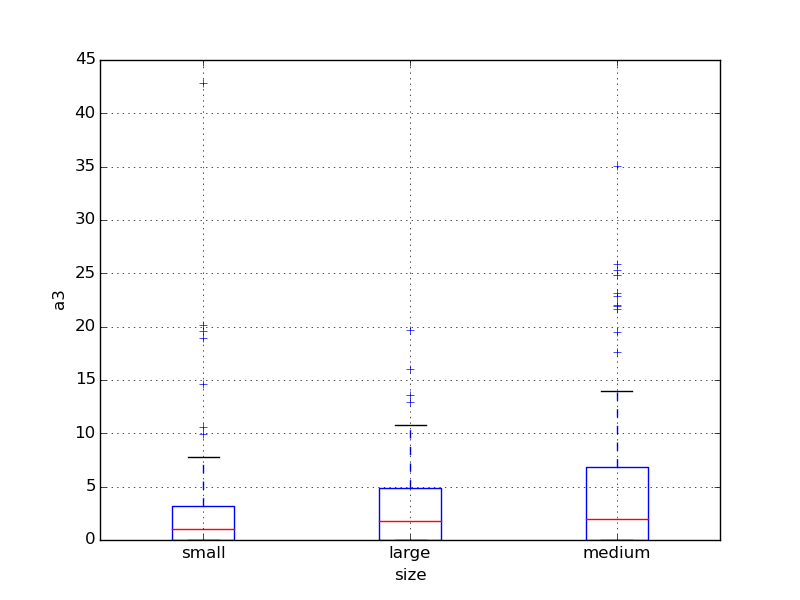
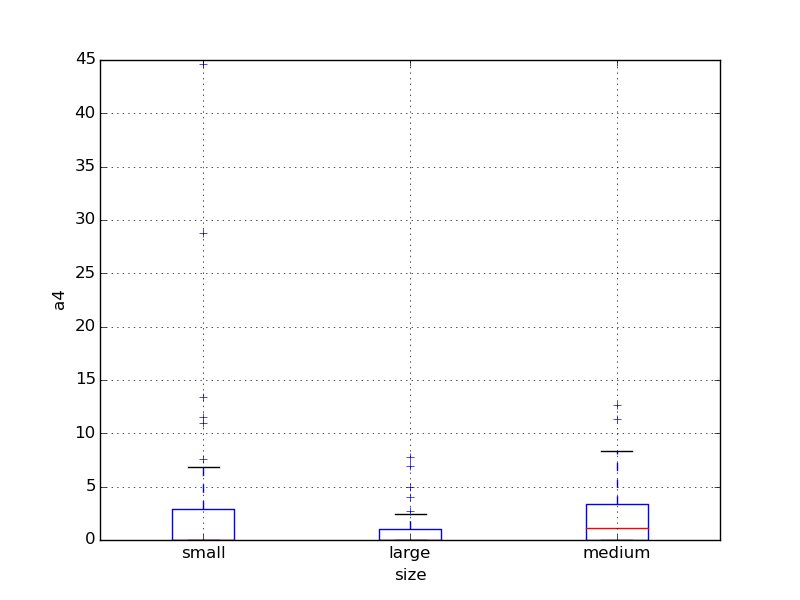
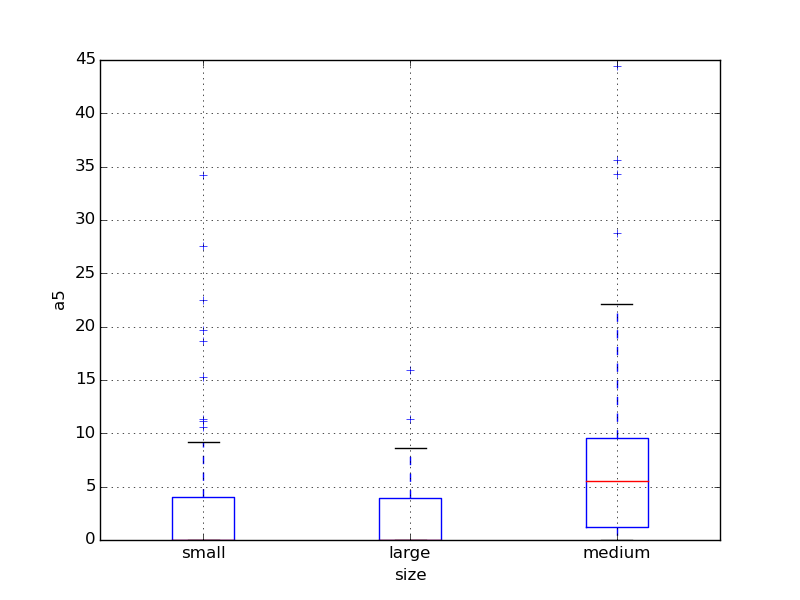
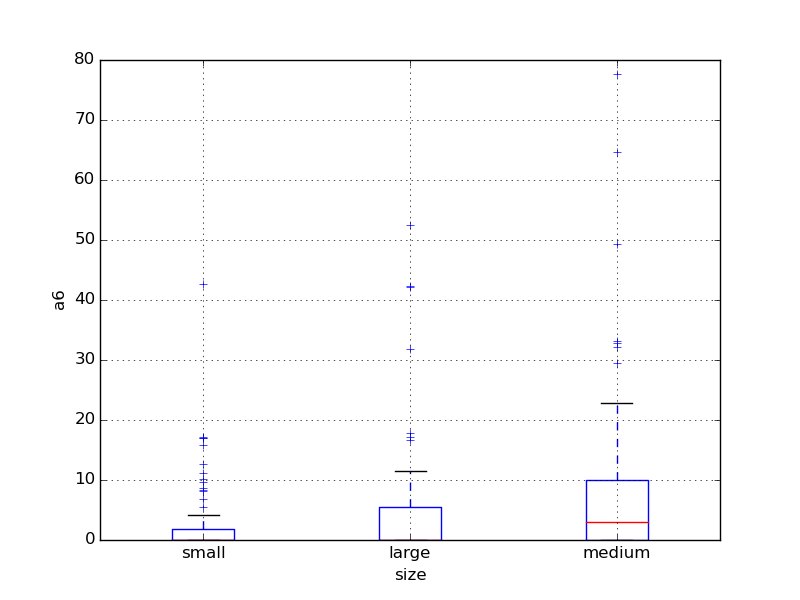
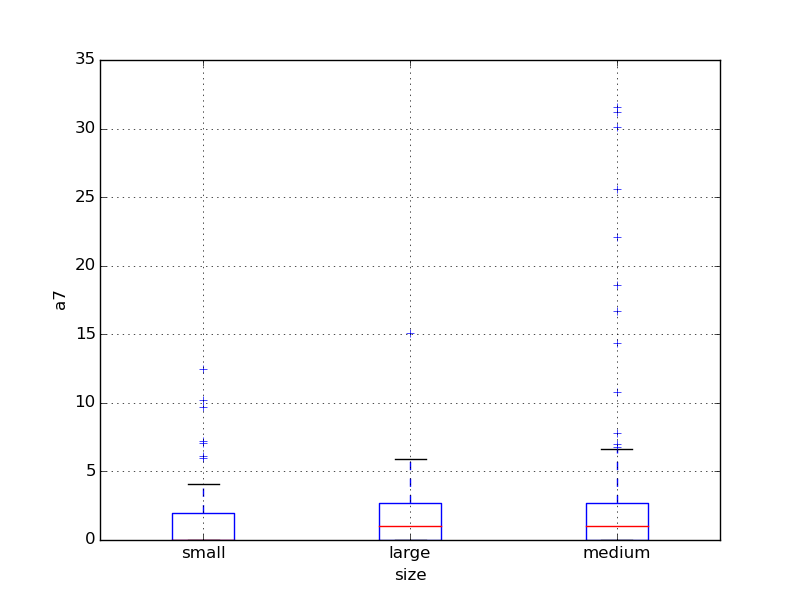


3、对7种海藻，其数量与标称变量的条件盒图

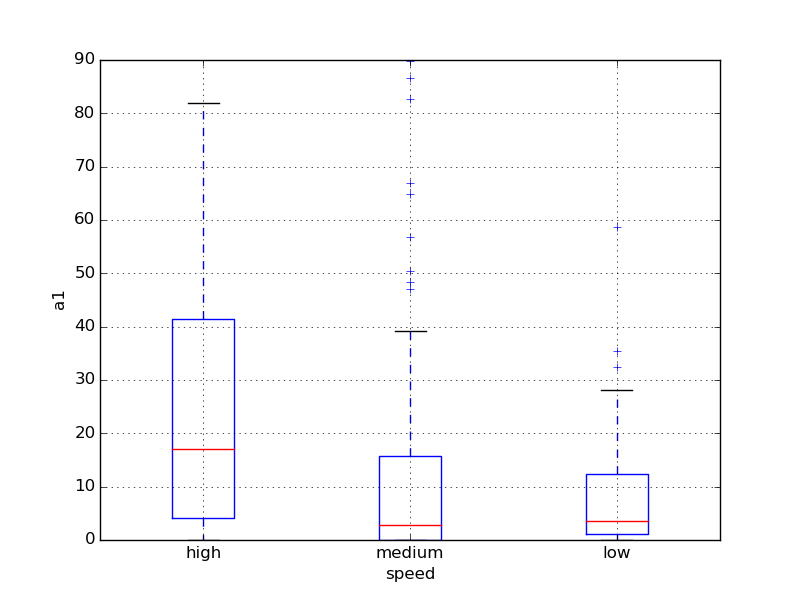
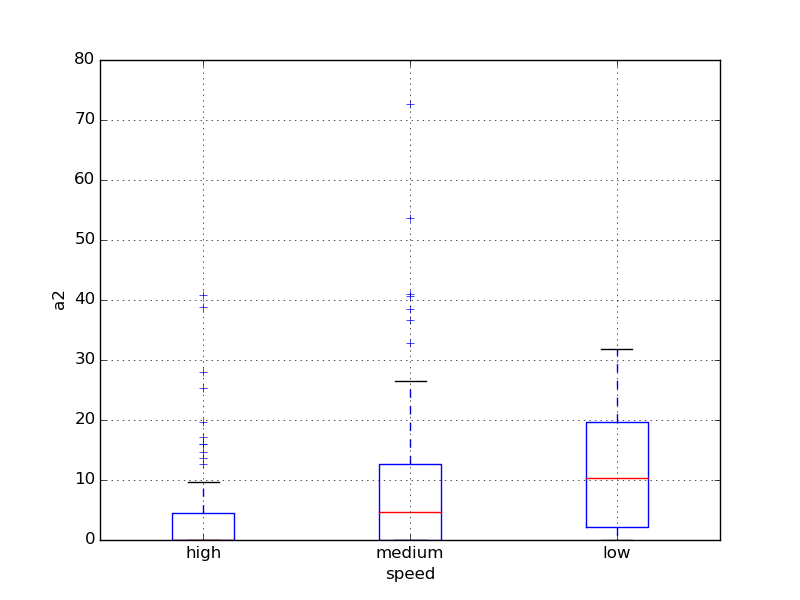
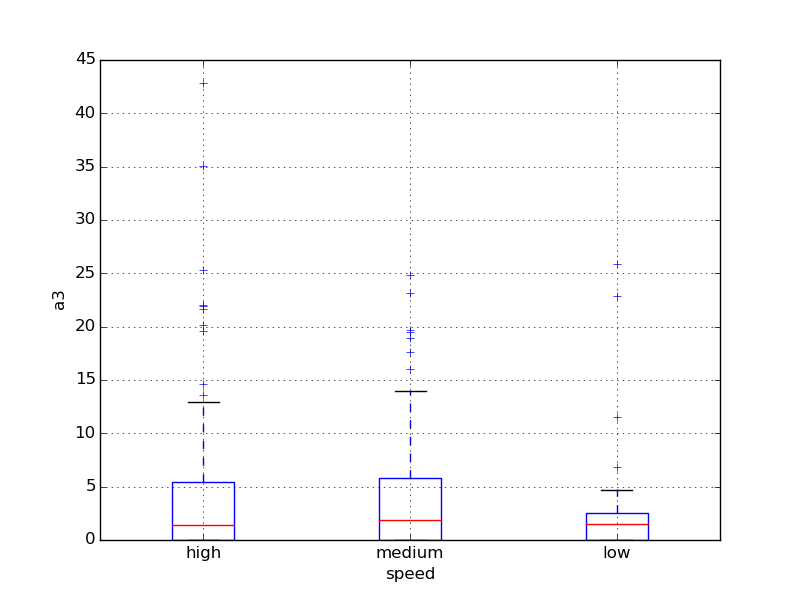
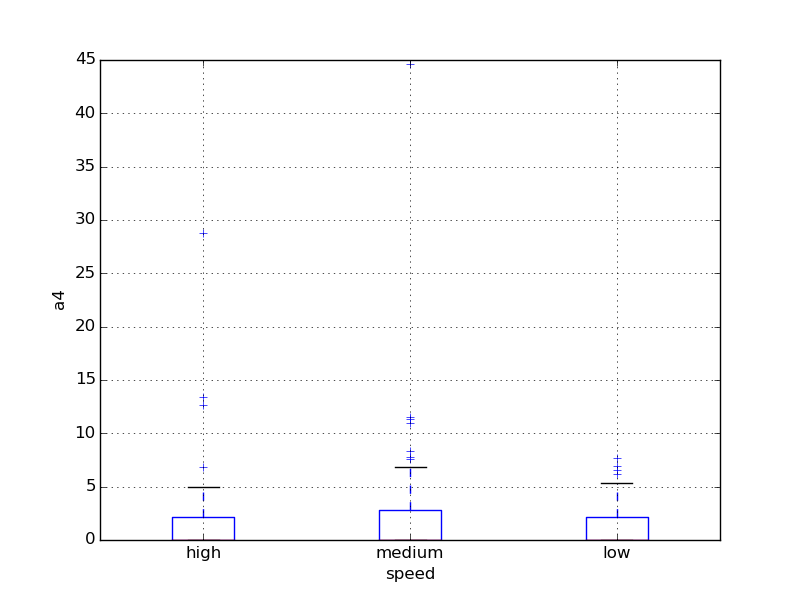
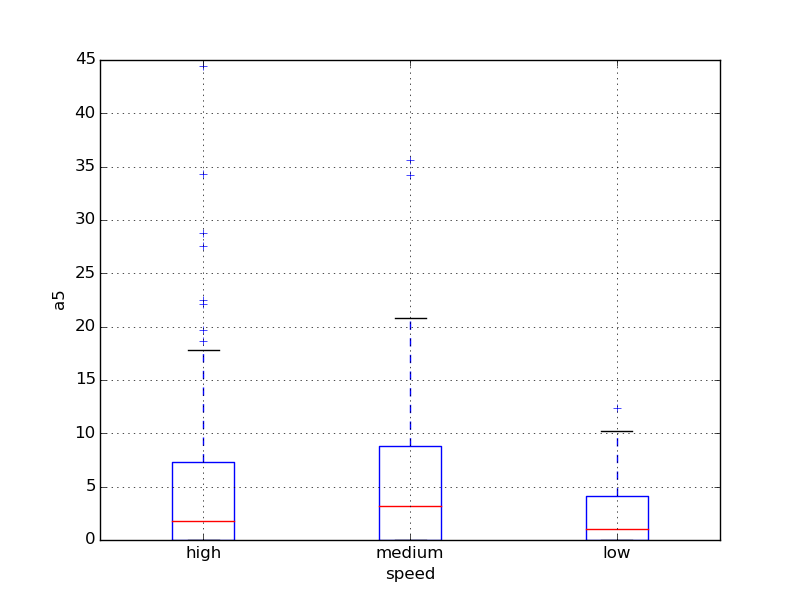
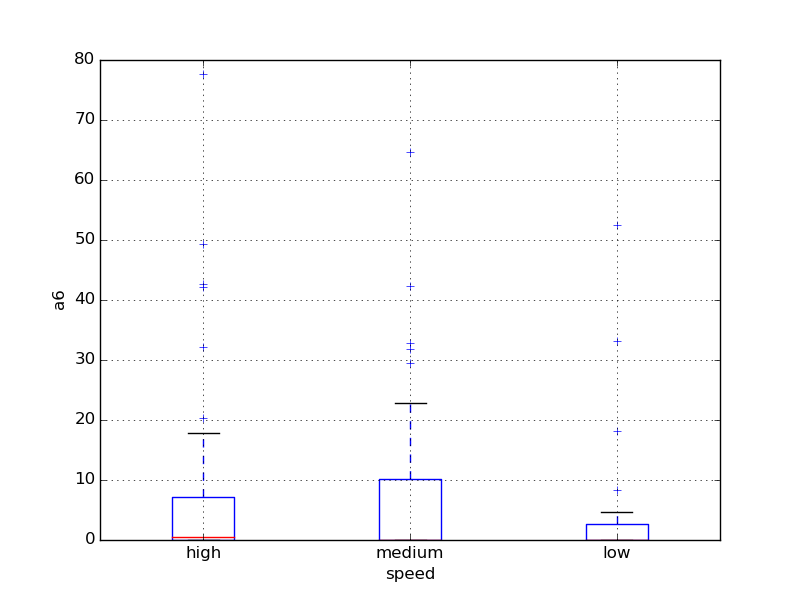
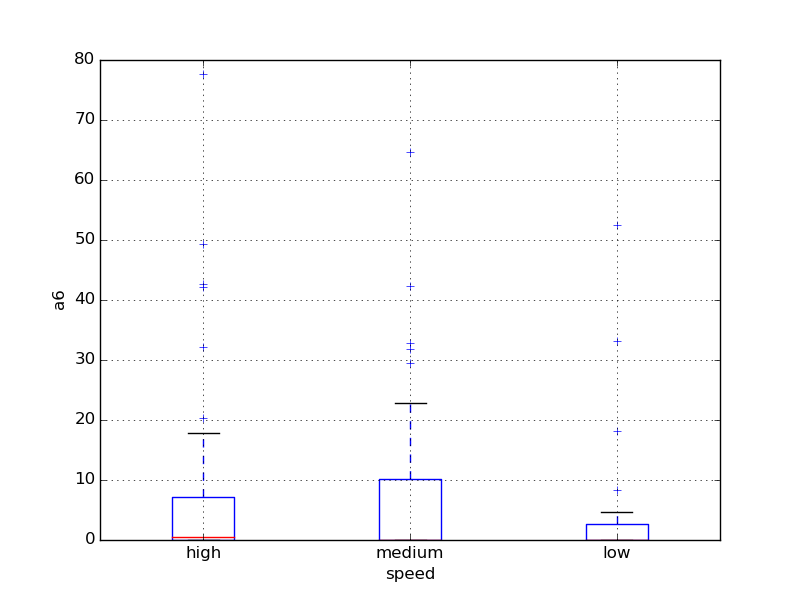
标称变量“季节”：



标称变量“河流大小”：

标称变量“河流速度”：

三、数据缺失的处理

1、将缺失部分剔除

参见Analysis\_without\_lost.txt

形如：

|winter|small|medium|8.00000|9.80000|60.80000|6.23800|578.00000|105.00000|170.00000|50.00000|0.00000|0.00000|0.00000|0.00000|34.20000|8.30000|0.00000|

|spring|small|medium|8.35000|8.00000|57.75000|1.28800|370.00000|428.75000|558.75000|1.30000|1.40000|7.60000|4.80000|1.90000|6.70000|0.00000|2.10000|

|autumn|small|medium|8.10000|11.40000|40.02000|5.33000|346.66699|125.66700|187.05701|15.60000|3.30000|53.60000|1.90000|0.00000|0.00000|0.00000|9.70000|

|spring|small|medium|8.07000|4.80000|77.36400|2.30200|98.18200|61.18200|138.70000|1.40000|3.10000|41.00000|18.90000|0.00000|1.40000|0.00000|1.40000|

|autumn|small|medium|8.06000|9.00000|55.35000|10.41600|233.70000|58.22200|97.58000|10.50000|9.20000|2.90000|7.50000|0.00000|7.50000|4.10000|1.00000|

|winter|small|high|8.25000|13.10000|65.75000|9.24800|430.00000|18.25000|56.66700|28.40000|15.10000|14.60000|1.40000|0.00000|22.50000|12.60000|2.90000|

|summer|small|high|8.15000|10.30000|73.25000|1.53500|110.00000|61.25000|111.75000|3.20000|2.40000|1.20000|3.20000|3.90000|5.80000|6.80000|0.00000|

|autumn|small|high|8.05000|10.60000|59.06700|4.99000|205.66701|44.66700|77.43400|6.90000|18.20000|1.60000|0.00000|0.00000|5.50000|8.70000|0.00000|

|winter|small|medium|8.70000|3.40000|21.95000|0.88600|102.75000|36.30000|71.00000|5.54400|25.40000|5.40000|2.50000|0.00000|0.00000|0.00000|0.00000|

2、用最高频率值来填补缺失值

参见Analysis\_freq.txt

形如：

|winter|small|medium|8.00000|9.80000|60.80000|6.23800|578.00000|105.00000|170.00000|50.00000|0.00000|0.00000|0.00000|0.00000|34.20000|8.30000|0.00000|

|spring|small|medium|8.35000|8.00000|57.75000|1.28800|370.00000|428.75000|558.75000|1.30000|1.40000|7.60000|4.80000|1.90000|6.70000|0.00000|2.10000|

|autumn|small|medium|8.10000|11.40000|40.02000|5.33000|346.66699|125.66700|187.05701|15.60000|3.30000|53.60000|1.90000|0.00000|0.00000|0.00000|9.70000|

|spring|small|medium|8.07000|4.80000|77.36400|2.30200|98.18200|61.18200|138.70000|1.40000|3.10000|41.00000|18.90000|0.00000|1.40000|0.00000|1.40000|

|autumn|small|medium|8.06000|9.00000|55.35000|10.41600|233.70000|58.22200|97.58000|10.50000|9.20000|2.90000|7.50000|0.00000|7.50000|4.10000|1.00000|

|winter|small|high|8.25000|13.10000|65.75000|9.24800|430.00000|18.25000|56.66700|28.40000|15.10000|14.60000|1.40000|0.00000|22.50000|12.60000|2.90000|

|summer|small|high|8.15000|10.30000|73.25000|1.53500|110.00000|61.25000|111.75000|3.20000|2.40000|1.20000|3.20000|3.90000|5.80000|6.80000|0.00000|

|autumn|small|high|8.05000|10.60000|59.06700|4.99000|205.66701|44.66700|77.43400|6.90000|18.20000|1.60000|0.00000|0.00000|5.50000|8.70000|0.00000|

|winter|small|medium|8.70000|3.40000|21.95000|0.88600|102.75000|36.30000|71.00000|5.54400|25.40000|5.40000|2.50000|0.00000|0.00000|0.00000|0.00000|

|winter|small|high|7.93000|9.90000|8.00000|1.39000|5.80000|27.25000|46.60000|0.80000|17.00000|0.00000|0.00000|2.90000|0.00000|0.00000|1.70000|

新数据与原始数据的比对图如下：

