L1 - Laboratório 1

Este documento apresenta a resolução do exercício de Laboratório 1 quem tem por objetivo apresentar e desenvolver técnicas estatísticas para auxiliar no estudo da ciência dos dados e extrair conhecimento do dados.

Perguntas:

a) qual foi a ferramenta/linguagem escolhida?
b) breve relato do seu processo de instalação/familiarização com a nova ferramenta, incluindo dicas para quem vai usá-la pela primeira vez;
c) descrição detalhada da análise realizada, incluindo ilustrações e código se julgar necessário;
d) sua conclusão após a realização da análise, revisitando a última pergunta da parte 1: os quatro conjuntos de dados correspondem ao mesmo fenômeno?

Respostas:

a) Dentre as ferramentas apresentadas no enunciado do exercício, por ter conhecimento em python, optei por uma das outras ferramentas. Já havia ouvido falar do Orange e feito um exemplo básico e por isso resolvi explorar melhor a ferramenta em um exemplo mais prático.

Entretanto, além do uso da ferramenta, montei também um jupyter notebook com quase mesma análise realizada aqui no software Orange.

b) Por ter familiaridade com python a instalação da ferramenta me pareceu ser bem simples. Optei pela instalação via pip (outro modo como a ferramenta pode ser instalada). Infelizmente para alguém que tenha pouco conhecimento no ambiente da linguagem python a instalação pode ser complicada.

Segui os passos conforme apresentados na página de "Download" da ferramenta, executando os comandos do programa pip (com uma diferença, neste caso, como mencionei,

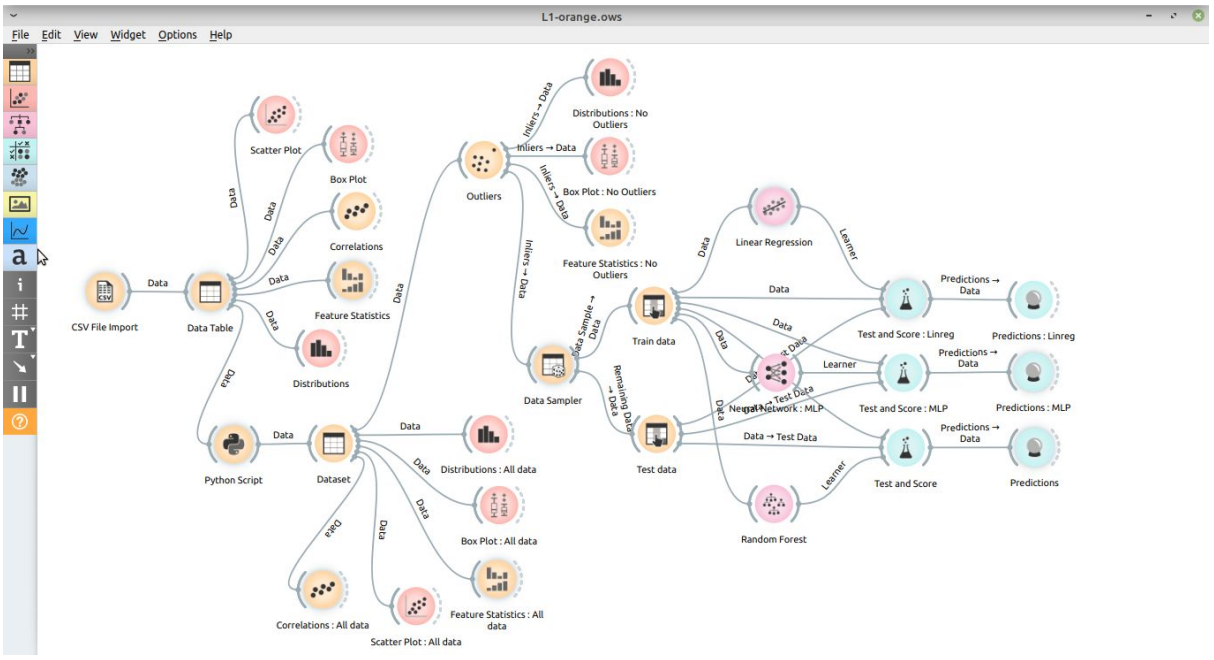
por conhecer o ecossistema da linguagem python, executei os comandos dentro de um "Virtual Environment").

Apesar disso, tive alguns problemas pois apesar dos passos no site da ferramenta mencionarem apenas a necessidade de instalação do Orange, tive ainda que instalar a biblioteca PyQt5 e PyQtWebEngine via pip também (ambos pacotes necessários para abrir a ferramenta e exportar os relatórios no formato pdf).

Passada a instalação inicial, segui os tutoriais no YouTube e explorei a página de Widgets da ferramenta onde existe alguma documentação de como usar as features dela.

O uso da ferramenta é bem simples e facilitado, algo ruim, foi que caso quisesse gerar estatísticas específicas apenas para o par de variáveis (x_i / y_i) teria de trabalhar os dados dentro da ferramenta (totalmente possível). Outra questão foi a geração deste relatório com os dados, a exportação dos gráficos e dados apesar de simples acaba gerando grande trabalho de salvar / copiar / colar.

Por fim, abaixo apresenta-se o diagrama completo criado dentro da ferramenta Orange para este exercício de laboratório.



Pipeline completo criado para o exercício de laboratório 1.

c) Conforme solicitado anteriormente via formulário, realizei inicialmente via Orange os mesmos cálculos solicitados.

No caso da ferramenta, ela possui um componente ("*Feature Statistics*") que facilita em muito a geração das estatísticas básicas dos dados conforme apresentado na Figura 1.

Figura 1. Estatísticas básicas.

Assim como nos cálculos realizados manualmente, podemos verificar que a Média e o Coeficiente de Variação (dados pelas colunas Center e Dispersion) são iguais para todos os x_i (ou seja média 9.00 e coeficiente de 0.35) e para os y_i (média 7.50 e coeficiente de 0.26). Com estas informações pode-se dizer que os dados fazem parte de um mesmo fenômeno, mas nesta tela ainda é possível observar também a distribuição dos valores, permitindo verificar que apesar da média e ser a mesma a distribuição (principalmente dos valores de y) são diferentes.

Em outro componente (“Correlation”) pode-se visualizar a correlação entre todos as variáveis, e até mesmo entre elas (como é possível notar na Figura 2), permitindo notar uma correlação linear positiva (principalmente entre os pares x_i e y_i com valores iguais a 0.816).

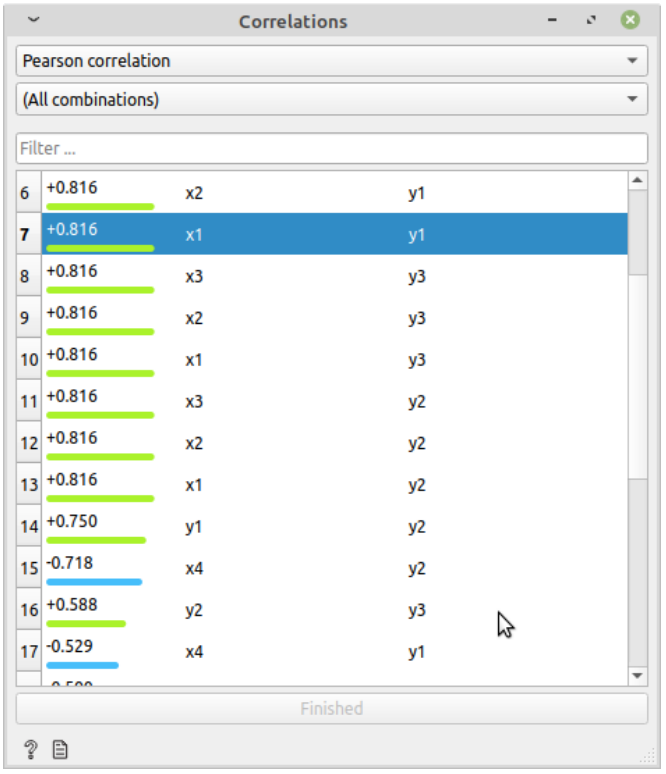


Figura 2. Correlações entre as variáveis.

A correlação pode ser visualizada apresentando em gráficos as variáveis em seus eixos, podendo ser percebida em 3 dos 4 casos uma linearidade crescente entre x e y. Conforme é possível notar na Figura 3 e mencionado anteriormente, tirando o par x_4/ y_4 todos os outros conjuntos de variáveis possuem uma tendência de crescimento linear positivo (principalmente x_1 e x_3, x_2 nem tanto).

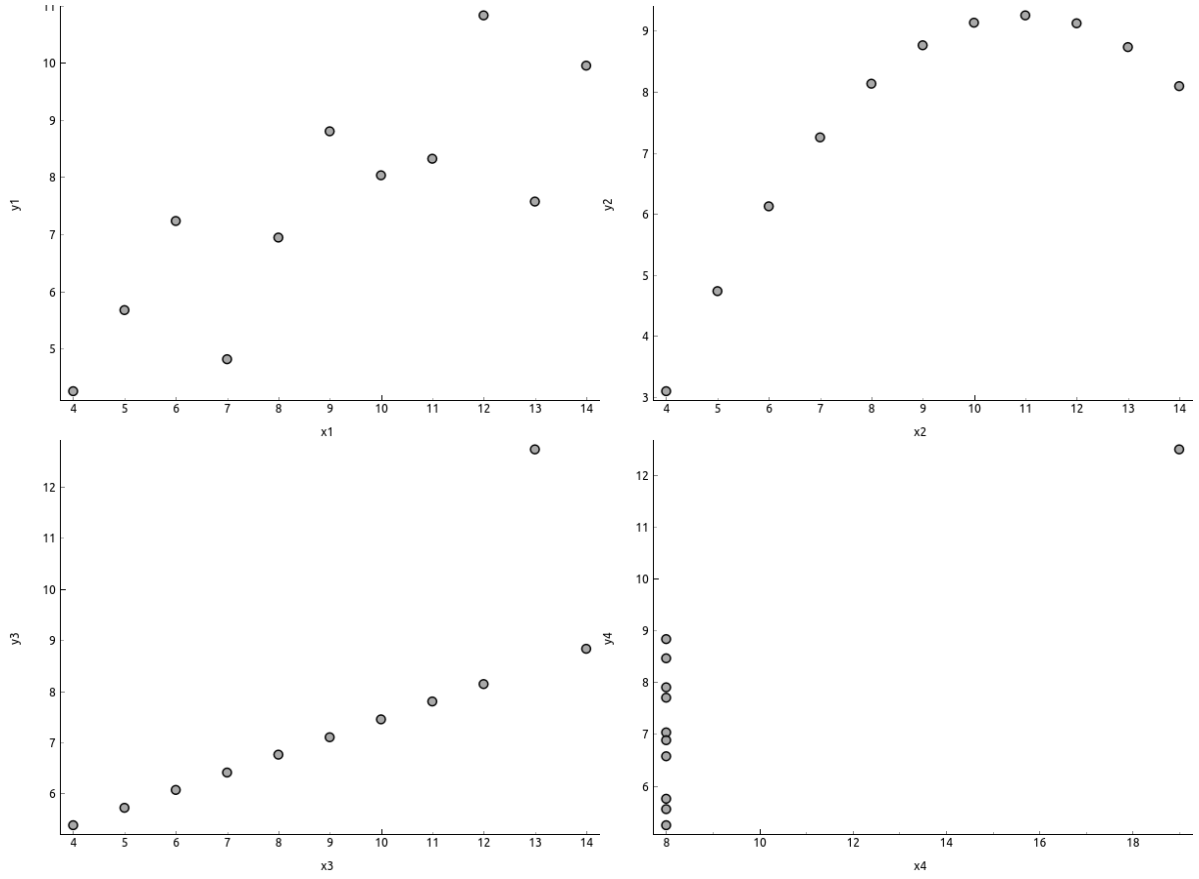


Figura 3. Gráficos de pontos para visualizar a possível correlação de x_i com y_i.

Além do apresentado acima, para visualizar melhor a distribuição dos valores das variáveis (que podem ser visualizados de modo parcial na Figura 1), dentro do software Orange, é possível usar o componente "*Distributions*". Com este pode-se gerar e apresentar gráficos das distribuições das variáveis e até mesmo apresentar alguma distribuição estatística dentro dados. Por exemplo, para os dados de y, pode-se apresentar a curva de uma distribuição normal (gaussiana) (Figura 4), entretanto a que parece ser melhor representada pelos valores das variáveis foi uma distribuição de estimativa de densidade de kernel (Figura 5).

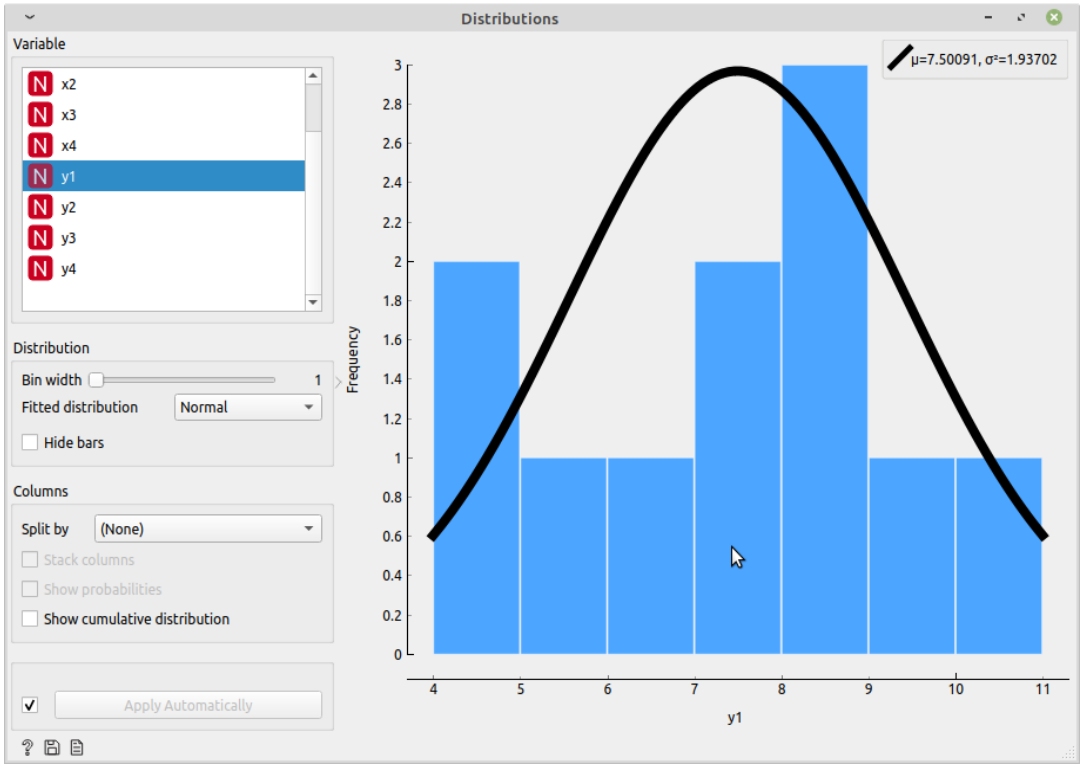


Figura 4. Possível distribuição gaussiana de y1.

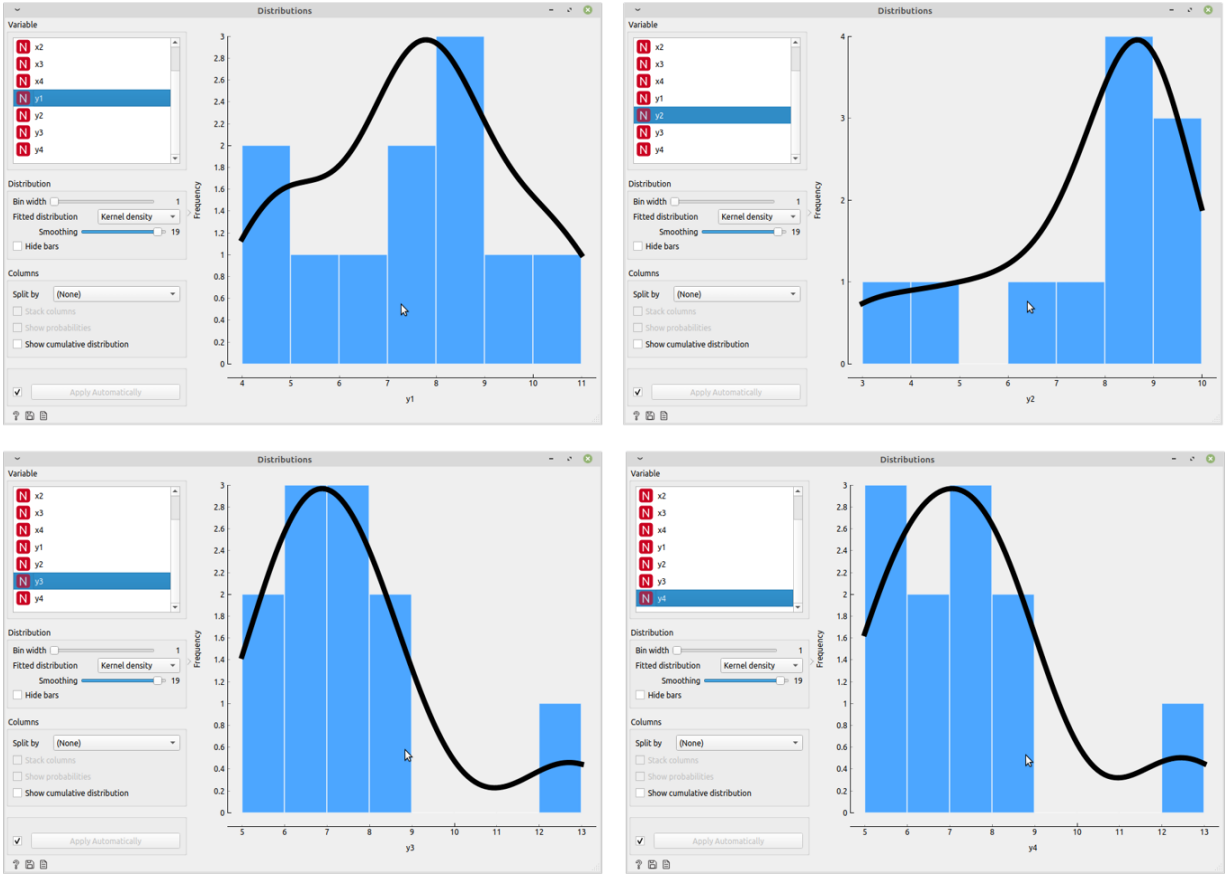


Figura 5. Estimativa de densidade de kernel para cada y valor.

Este fato da distribuição de estimativa de densidade de kernel ser melhor representada pelos dados, fica bem mais visível ao colocar todos os valores de x e y em um único conjunto sequencial de elementos (todos os valores de x foram concatenados sequencialmente, feito o mesmo para y, dessa maneira tem-se um único conjunto com 44 valores de x e y, ao invés dos 4 conjuntos separados), em comparação com uma distribuição normal (conforme Figura 6).

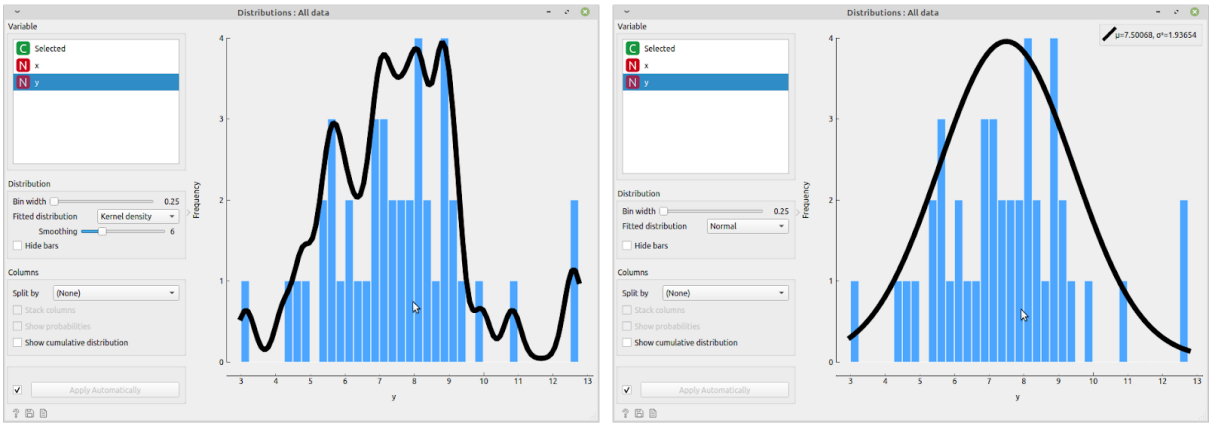


Figura 6. Estimativa de densidade de kernel (esquerda) e gaussiana (direita) para todos os y valores.

Entretanto esta questão (a definição do uso da distribuição de estimativa de densidade de kernel), deve ser mais analisada a fundo, pois é possível notar a existência de alguns valores de y que podem ser “Outliers” (principalmente nos pares x_3 / y_3 e x_4 / y_4). Aplica-se então componente da ferramenta que faz a identificação de “Outliers”, por padrão utiliza-se a estimação de covariância para a remoção dos possíveis “Outliers”.

Realizando tal operação, são eliminadas amostras (conforme Figura 6) que estão nas periferias da distribuição. Portanto, dessa maneira, temos uma distribuição sem os “Outliers” identificados e filtrados, permitindo que a distribuição gaussiana molde-se melhor as amostras existentes (Figura 7). Essa remoção de “Outliers” não segue a mesma regra se utilizado para cada um dos conjuntos em separado, variando o tipo de algoritmo a ser utilizado na identificação dos “Outliers”.

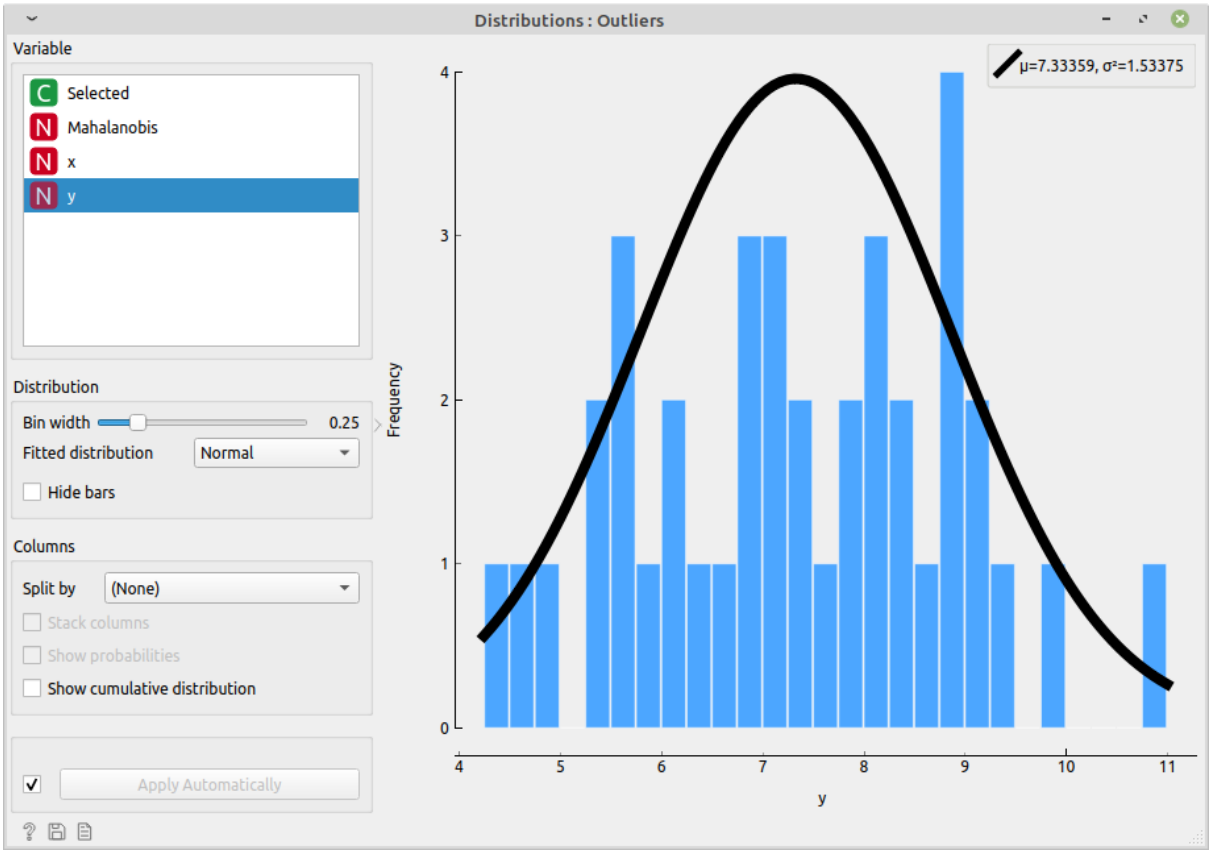


Figura 7. Distribuição Gaussiana para todos os y valores.

Entretanto através da comparação dos diagramas de caixa (Figura 8), é possível observar que a remoção destas amostras definidas como “Outliers” faz com que a mediana da distribuição e a média não estejam mais próximas uma da outra, inferindo portanto que a nova distribuição não seria tão simétrica quanto a anterior.

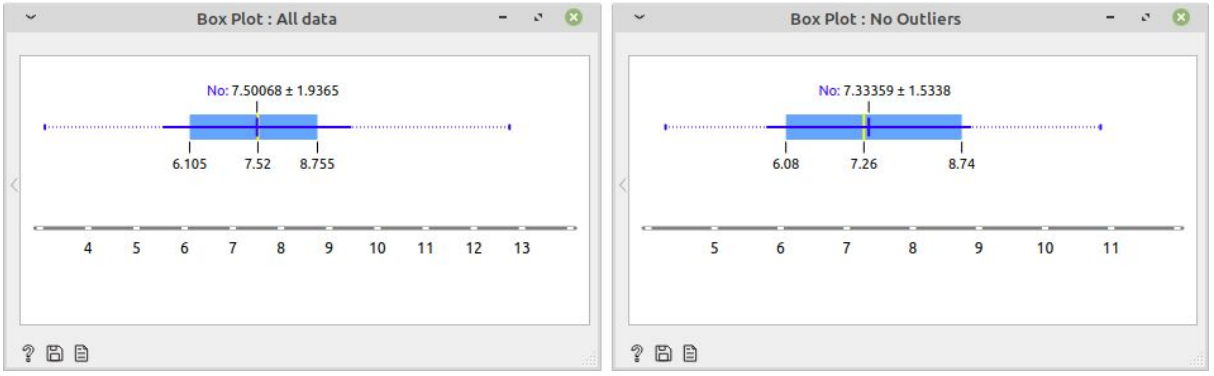


Figura 8. Diagramas de Caixa. A esquerda o diagrama com todos os valores de pares x, y de todas as amostras, e a direita os diagrama após a remoção de possíveis “Outliers”.

Portanto a remoção dos “Outliers” apesar de aproximarem melhor o perfil das amostras em direção a uma gaussiana, os gráficos de diagrama de caixa apresentam um desvio informando que a nova distribuição não seria simétrica, mas teria uma assimetria positiva (devido a mediana aproximar-se para o primeiro quartil).

d) Revisitando a última pergunta da parte 1 e após análise acima, é possível afirmar que os 4 experimentos fazem parte do mesmo fenômeno, observando apenas as estatísticas básicas (Média, Desvio Padrão e Correlação), entretanto ao visualizar os pontos nos gráficos e verificar a possível distribuição dos valores e mesmo com a possível remoção de “Outliers” do conjunto completo dos dados, pode-se observar que não é possível afirmar com 100% de certeza que ambos os 4 conjuntos fazem parte de um mesmo fenômeno ou mesmo que podem ser representados por uma distribuição normal.