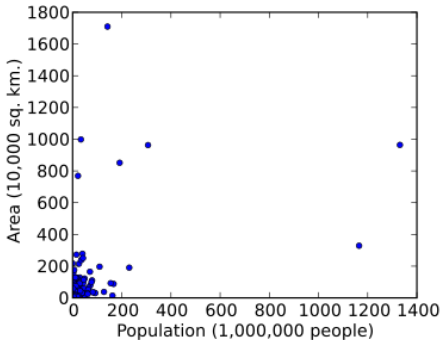
01. Self-assessment questions

January 5, 2022

- Look at the scatterplot below[1]



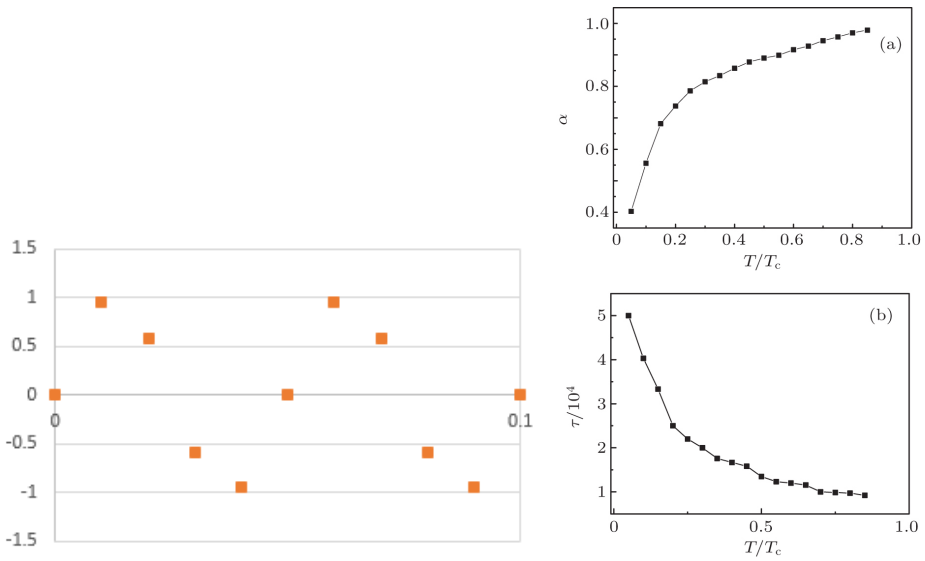
  Can you clearly see if there is a dependence between Population and Area? Are there any transformations that can help to unveil the possible dependency? If yes, would you apply such transformation on the x-axis, y-axis or both?

- Look at the three plots below [2] [3]

[1] Attribution Skbkekas, CC BY-SA 3.0 <http://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons

[2] From `http://cpb.iphy.ac.cn/article/2015/cpb_24_7_070501/cpb142814f4_hr.jpg.html`

[3] From P. Scheidler "Understanding the Basics of Fourier Transforms"



For each graph, tell if there is a dependence between the x-axis and y-axis variables. If yes, would this result in a null, positive or negative Pearson's correlation coefficient?

- Assume one of your features has 4 values, you can encode them in two ways: (i) Transform them in 0,1,2,3 or (ii) Apply One-Hot Encoding. When would you choose (i) and when (ii)?
- Suppose you have a dataset with 200 columns. One column has 0.1% of missing values. Another column has 30% of missing values. What pre-processing would you apply before applying any supervised or unsupervised machine learning methods? [See the code related to pre-processing]
- Suppose you have a column with the following values: student, medical doctor, engineer, cashier, no-info-available. How would you transform such column in numerical form?
- Suppose you have a column with the following values: very bad, bad, medium, good, very good, excellent, no-opinion. How would you transform it in numerical form?
- Write the formulas of variance (of a feature), standard deviation, Pearson's correlation coefficient.
- Explain how you construct a boxplot and what is the meaning of all the lines that compose it.
- Explain how to calculate the 35-th percentile.

.