CASE STUDY 012
[Machine Learning: R]
Crime prediction

Difficulty Level: 3 of 3

Disclaimer: SuperDataScience has no affiliation with data sources. The scenario is made up for educational purposes.

You are a Data Analyst working for the police department. You have been given a data set with a large number of variables and have been asked by the police chief to build a regression model to predict crime rates.

1. Clean variable_names and set the names of data to these clean names
2. The police chief is only interested in the following variables so create a new data frame with just these variables included: 'ViolentCrimesPerPop', 'pctUrban', 'agePct16t24', 'PctUnemployed', 'medIncome'
3. Check for correlations between medIncome and PctUnemployed. Plot these variables to confirm correlations
4. Split the data into training and testing sets using set.seed(123)
5. Build a linear regression model using pctUrban, agePct16t24 and whichever of the variables from question 3 is best correlated with ViolentCrimesPerPop
6. Predict the ViolentCrimesPerPop for the testing set
7. Calculate the R^2 of the testing set

Good luck!

Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file. Data source: Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science