# **Feature Selection and Feature Engineering**

# Introduction

The goal of data science as a field is to produce insights from data and when it is practiced effectively it should produce actionable insights. The features selected for supervised and unsupervised learning algorithms should have strong relationships to the primary outcome of interest.

The target variable in the Decision Tree Assignment is the rate of violent crime. For this assignment you will use the Vital Signs data and transform a variable or variables to create a new variable that you believe will have some predictive power for the rate of violent crime. In the first part of the assignment you will discuss the way in which the feature you are creating is was informed by the assigned readings. For the second part of the assignment you will transform the data in R.

The assignment will be graded on two criteria. The first is the degree to which you read, absorbed, and integrated the readings into your thinking and the quality of your communication about that. The second is the thought and effort you put into constructing the feature – how you assessed the raw data, the transformations, recoding, restructuring, or reformulation of the original data that went into your process. For this assignment, the grade will depend on 1) your thinking and 2) your effort in constructing a new variable.

For the Decision Tree, you will test the predictive capacity of the feature you created. So, you will be putting your work to the test. Your final target variable for that assignment will be rates of violent crime for 2017.

The most important reading for this assignment is "Predicting Gentrification Using Longitudinal Census Data." The analysis in this article demonstrates innovative data transformations grounded in subject matter expertise. The value of "How Israel Spots Lone-Wolf Attackers" is the example it presents of both thinking outside the box and also identifying a feature that is genuinely effective in its predictive capacity. And "How Can Baltimore City Prevent Housing Abandonment?" is a good discussion of the broad search for effective features.

The Blackboard section "Feature Selection and Feature Engineering" in Lesson 6 has a long list of types of feature engineering. Use these as a source of ideas. The methodology for some of the transformations on the list is beyond the scope of the course, but others can be used for this assignment.

## Directions

Use the Vital Signs 2010 data from Assignment02 for the assignment. You may also use any or all of the other years for the Vital Signs data.

Use the assigned readings from Lesson 6 as your guide for how to understand feature selection and feature engineering. Use *Practical Data Science with R*, Chapter 4, Section 4.1, and *R in Action*, Part 1, Chapter 4, for information on how to make transformations to data.

The completed assignment should adhere to the following guidelines:

* 1. Include your answers on the assignment document.
  2. Write your answers using complete sentences with correct punctuation, grammar, and spelling.
  3. Submit your completed assignment through the Blackboard portal in Lesson 6. It may be submitted early, if desired.

1. What aspect of the methodology or reasoning in the assigned readings figured into your idea for a feature? Describe the feature you plan to engineer. Include a discussion of the original variable or variables in your description. Explain the reasoning for your choices. *(5 points)*

Predicting gentrification reading – from that reading I plan to visualize, at least attempt to, visualize my results per neighborhood (minus the GIS sections).

How Can Baltimore City reading – from that reading I got the idea to use multiple secondary variables to try to predict another variable that I do not have readily available.

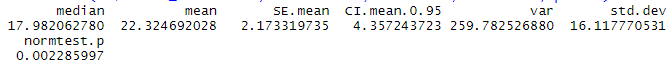
How Israel Spots reading - From that reading I ran into a wall, sadly. I know it’s possible to use a web crawler to explore social media accounts, but I do not have the computer power to extract and store data from all users in Baltimore city (nor the know-how in R). It’s exciting to think about the possible use-cases none-the-less.

I plan to engineer a variable that calculates the difference of income inequality in each neighborhood. The original variables I plan to use are, “hh25inc17” and “hhm7517”. I believe this to be a strong predictor of violent crime as it can point to wealth distribution (or lack of) in a neighborhood.

1. Explain how you transformed the data and include the code that you used below your answer. Run the stat.desc function from **pastecs** and the skim function from **skimr** to run descriptive statistics for your new variable. Paste the results below. *(5 points)*

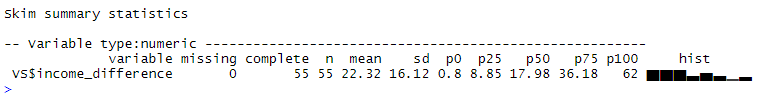
I transformed my data by introducing a new column to the dataset, “income\_difference”. I populated this column from the absolute difference in the percentage of households earning more than $75,000 minus > $25,000.

**pastecs** -





**skimr** -



While not explicitly requested, nevertheless here’s my code for the entire file:

#import dataset

VS <- read.csv("C:/Users/carlb/Desktop/Week 6/VS.csv", header=TRUE)

#convert to df

VS<-as.data.frame(VS)

#generate feature w/blanks

VS["income\_difference"] <- ''

#calculate difference of income (in absolute values)

VS$income\_difference <- abs(VS$hhm7517 - VS$hh25inc17)

#load the lastecs package

require(pastecs)

#run new feature though pastecs

stat.desc(VS$income\_difference, basic=FALSE, desc=TRUE, norm=TRUE, p=0.95)

#load the skimr package

require(skimr)

#run new feature though skimr

skim(VS$income\_difference)