**Data Wrangling Challenge**

**Problem Statement**

U.S. Department of Education

Policy Priority: The Department of Education is focused on ensuring that parents, students, and policymakers are able to use its publicly available data to take meaningful action to improve outcomes.

1. Supporting Decision-Making for Student Sub-populations and their Families  
   Problem: Certain mobile or disconnected student populations entering or reentering the community could greatly benefit from data and resources to support their wellbeing and success. Such students and families often lack information that is necessary to distinguish between their school options, access services, and identify affordable housing near high-quality school and in safe neighborhoods that have access to transit and employment.

**Use Case**

Maria is a 25-year-old US Army veteran, newly returned to the civilian workforce. She has recently completed a six-year commitment with the Army. During her time in the Army, she worked in supply management and logistics. She has decided to pursue a degree in Management Systems and Information Technology.

Maria has asked you to use your skill with data to help her search for the best school for her. She is willing to relocate anywhere in the continental United States, but she has a few criteria that her ideal schools must satisfy: 1) safety (low crime), 2) urban -- Maria wants to live the big city life, and 3) start-ups -- the school should be in a metropolitan area that ranks highly in entrepreneurialism (she plans to find an internship at a startup while she studies).

Maria would like you to help her narrow down her search to a list of schools that she can investigate more closely before making her decision.

**Your Task**

1. Produce a dataset of schools which satisfy all of Maria's criteria
2. Rank them from best to worst according to the same criteria.

Maria's schools must:

1. be in an urban/metropolitan area.
2. be in a city that ranks 75th percentile or higher on Kauffman's start-up rankings.
3. be below 50th percentile in overall crime.
4. offer a 2-year or 4-year degree in Information Technology/Science.

All the data you need is provided with the exception of the entrepreneurial data, which can be accessed at <http://www.kauffman.org/microsites/kauffman-index/about/archive/kauffman-index-of-entrepreneurial-activity-data-files> (Metro Area Components Data 2015)

Tips:

1. Read the data dictionaries or codebooks to figure out what the variables mean and which ones you will need to use.
2. Eliminate unneeded columns.
3. Look for suitable columns to join the tables on.
4. Perform any cleaning and standardization needed to facilitate the joins.
5. Engineer a summary variable for school crime so that we can compare schools by levels of crime overall.
6. Eliminate from the data all the data points that fail to satisfy Maria's criteria.
7. Engineer a method for ranking the schools in consideration of all of Maria's criteria taken together.

**Extra Credit**

1. Maria doesn't like the cold. Find and integrate temperature data. Eliminate any schools located in cities/areas below the 25th percentile in average temperature.
2. Replicate this process using R or SAS.

* *This was created as a training exercise in data wrangling/preparation.*
* *Prerequisites include scripting skills with Python Pandas or R (or another data manipulation tool) to perform data cleaning, transformations, joins, etc.*
* *No additional subject matter or data analysis knowledge is required.*
* *The task is a specific and hypothetical use case based on a problem statement proposed by the Department of Education in Opportunity Project documentation.*
* *The participant who completes the exercise will demonstrate the ability to clean, join, and organize data from disparate tables and formats in the process of solving a realistic problem.*

**Solutions**

* [<https://data.world/exercises/wrangling-exercise-1-solutions>]
* Submissions. Showcase your work and help others. We are accepting exemplary scripts, visualizations, and explanations to add to the exercises' solutions. Reach us on our [community Slack](https://slack.data.world/) or at exercises@data.world for submissions or inquiries about featuring your work here for posterity!