COM 307 Machine Learning/Data Mining Project 2: Wrongful conviction data (due by October 12 before class start)

In this project you will be using real-life data to see how different variables associate with wrongful convictions and predicting the likelihood of being exonerated for a crime based on several factors. In order to do this, you will be implementing a Naïve Bayes Classifier, as discussed in class (lecture 9). In order to do this, you will need two sources of data:

1. Information on exonerations for wrongful convictions. For this, we will be using the National Registry of Exonerations, hosted by the University of California at Irvine's Newkirk Center for Science & Society, the University of Michigan Law School, and the Michigan State University College of Law. It can be found online at https://www.law.umich.edu/special/exoneration/Pages/about.aspx, and the data itself can be found in table form at https://www.law.umich.edu/special/exoneration/Pages/detaillist.aspx or on Moodle. If you use the Moodle spreadsheet, please find the key for translating the tags on the website.

2. Information on convictions that are not exonerations. There aren't really databases for this, so we will be using total convictions. You can then find convictions that are not exonerated by subtracting the exonerations from the total convictions. This data is available from the United States Sentencing Commission. We will be using the data from 2018 because there's a lot of data and it's frankly too much to go through every year for a single project (could work as a final project, though!). It is available on Moodle, and there is a lot of it (it's a big file). You will be particularly interested in the data contained in the following fields:

   - "OFFGUIDE" – describes the offense (see key on A-5 in the codebook)
   - "DISTRICT" – tells the location of the offense (see key on A-3)
   - "MONRACE" – Convicted individual's self-reported race (key on 39)

   You will use "USSC_Public_Release_Codebook_FY99_FY18.pdf" from Moodle as the key. There are exonerations for crimes going back to 1956. In order to model this using only 2018 convictions, you'll have to make choices. Tell me about those choices (see more below)!

[Submit]

Submit as a zipped file the following files:

1) Your code (in whatever language you used)
2) The probabilities that you generated for your model
3) A not too official write-up of what you found during this exercise and what you did when you reached decision points. How did you deal with the fact that you had only 1 year of conviction data but decades of exoneration data, for example.