ECON 282, Professor Hogendorn

Final Project Part 1

Find or compile a dataset to use for this project, looking for one that is at least 200 rows tall and also relatively fat, i.e. has a large number of variables per observation. If you are interested in sports, team or player performance data should work well. Many large datasets are available at the websites listed below. Be warned that it may take some data wrangling to get them into R.

www.data.gov	www.census.gov
research.stlouisfed.org	data.worldbank.org
www.nsf.gov/statistics/data.cfm	datahub.io
aws.amazon.com/datasets	www.kdnuggets.com/datasets/
www.icpsr.umich.edu/icpsrweb/ICPSR/	netdatadirectory.org/

Also try azure.microsoft.com/en-us/documentation/articles/machine-learning-use-sample-datasets/

Don't forget about the possibility of joining data from multiple sources.

Please hand in the following:

1. Describe the dataset in words.

2. What is dependent variable you are interested in? Why?

3. Show the `str()` of your data and the `summary()` of some of the key variables that interest you.

Part 2 of the project will ask you to use data visualization, build models using theory, and build models using machine learning. Make sure to choose data that you look forward to working with and learning from.