3 US College Scorecard

This project is the first project that led me to using BayesDB, the AI assistant i've described earlier.
This dataset has been carefully chosen by me and Vikash. We needed to build an analysis where the historical ones were misleading the readers, inclined to emotional behaviors and definitely splitting the communiity with different point of views. The Education was ideal. Ironically enough, the College Scorecard dataset was published by the U.S. Department of Education in October 2015[1].

The Economist triggered discussion about this dataset in an article describing a ranking system based only on earnings after graduation[2]. While the vision of BayesDB, and of probabilistic programming languages in general, is to come as an AI-enhanced medium to create consensus on different fields where historically, researchers, scientists, domains experts, barely succeed in agreeing with each other, the analysis on the US Education Scorecard would give a good material to show how non experts of the domains (I am certainly not, besides the only input about the models are only mathematics wise and surely not involving cultural knowledge for instance) could compute logical and complete analysis of a specific field.

3.1 Problem Statement

This huge dataset is going to be a good field for BayesDB to showcase its ability to understand any dataset and extracting useful insights of the fiels.

· The issue is that traditional rankings of American colleges do not focus on many variables and base its core analysis on metrics related to graduates earnings. In the meantime, the challenge here is to be able to compute a transparent value-added for each college in order to quantify the salary boost the students are receiving from attending such schools.

· Also, current rankings prefer translating the opportunities given by a school, via its network, its partnerships... and not on the hard work and intelligence of its graduates.

· In this following document, we will ask BayesDB different questions in terms of relations between variables and distribution of some metrics to compare these results to our intuition and to other iterations of the same models.

As a result, we should expect our tool to take into consideration every relation between variables and show us intuitive dependencies and simulations.

We will organize the study in different parts:

[1] https://collegescorecard.ed.gov/data/
[2] http://www.economist.com/blogs/graphicdetail/2015/10/value-university

1. The first part will consist of analyzing a small subsample, given our initial selection of 50 variables, of 100 observations (colleges). We'll be plotting the dependencies heatmap and be adding more and more observations with fixed and variable number of models and iterations. Both cases would show interesting results.

2. Then, using gpmcc, a new implementation of CrossCat from the the lens of generative population model (GPM), we'll simulate distributions of a few variables and compare them with the existing distributions and our intuition. We'll try the same experiment by inferring missing values only.

3. On the same variables, we'll ask BayesDB to find unlikely data, whether they are outliers or input errors.

4. Finally, we'll simulate colleges by size, selectivity and students earnings after graduation and challenge stereotypes about colleges thanks to these results.

3.2 Dataset

The dataset published by the U.S. Department of Education is composed of 18 years (from 1996 to 2013) of data about more than 7,000 schools. More than 1,700 variables have been measured each year for each university. The number of universities fluctuates according to the creation of new schools and the closing of existing ones.

The overall dataset contains more than 215 million values with 43% of them missing.

Figure 1 highlights the number of variables measured each year. Obviously, we've been able to measure more and more variables through the years, but for some reason many data are missing in 2013 (the last year this study has been conducted). Obama's administration college database is way more transparent and better quality compared to datasets used by US rankings publishers. It has better data on family income, family education levels of entering students and more sophisticated measures of degree completion as well as loan repayment. These numbers have been generated in particular by matching students loans to their actual tax returns. As a result, we are able to compare professionals' earnings to their student characteristics.

Note: The data only includes students who applied for financial aid and thus is missing all the students with well-off parents. Also, as the earnings data only take into consideration 10 years after starting college, one could argue that this scope of time is too small to include future high earners, as they would still be students (e.g. Ph.D., post-doctorate).

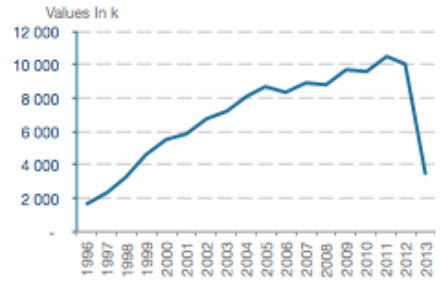

Figure 1: Data available per year

The number of schools from one year to another is also varying. Here is an overview of how many schools we have metrics of. One can tell that the overall trend is showing an increase in number of schools even though from 1996 to 1998 we can observe more schools closing than being created. One interesting result could be to characterize the fall of trends of these schools and be able to predict future outlook of current schools based on their most recent data.

```
bdball = bayeslite.bayesdb_open("dfall.bdb", builtin_metamodels=False)
bdbcontrib.barplot(bdball, '''
SELECT year ,
COUNT(OPEID) AS "Number of schools"
FROM dfall
GROUP BY year
ORDER BY year asc
''');
```

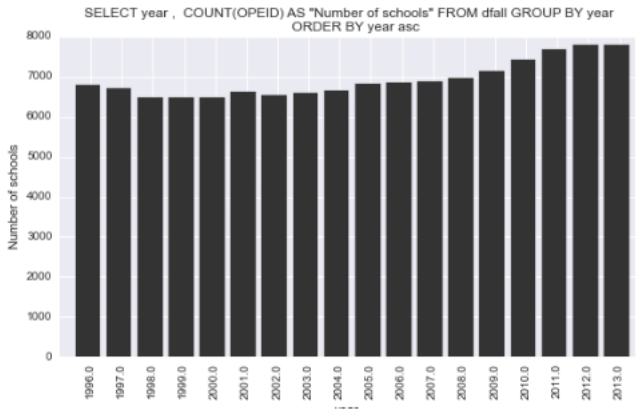

Figure 2: Number of colleges in the US per year

Our 57 variables include some financials and specs about the universities. Here is a full list of the variables selected for the study

var_name	description
Financials	
DEP_DEBT_MDN	The median debt for dependent students
LO_INC_DEATH_YR2_R	Percent of low income (between 0 and 30k in nominal family income) students who died within 2 years
MD_INC_DEATH_YR2_R	Percent of middle-income (between $30k and 75k in nominal family income) students who died within 2 years
HI_INC_DEATH_YR2_R	Percent of high income (more than 75k in nominal family income) students who died within 2 years
PELL_DEATH_YR2_RT	Percent of students who received a Pell Grant at the institution and who died within 2 years at original institution
NOPELL_DEATH_YR2_RT	Percent of students who never received a Pell Grant at the institution and who died within 2 years at original institution
LOAN_DEATH_YR2_RT	Percent of students who received a federal loan at the institution and who died within 2 years at original institution
NOLOAN_DEATH_YR2_RT	Percent of students who never received a federal loan at the institution and who died within 2 years at original institution
NOLOAN_ENRL_ORIG_YR	Percent of students who never received a federal loan at the institution and who were still enrolled at original institution within a year
DEATH_YR2_RT	Percent died within 2 years at original institution
LO_INC_COMP_ORIG_Y	Percent of female students who transferred to a 2-year institution and whose status is unknown within 8 years
MD_INC_COMP_ORIG_Y	Percent of middle-income (between $30k and 75k in nominal family income) students who died within a year
HI_INC_COMP_ORIG_Y	Percent of high-income (over in nominal family income) students who died within a year
PELL_COMP_ORIG_YR2	Percent of students who received a Pell Grant at the institution and who completed in 2 years at original
NOPELL_COMP_ORIG_YR2	Percent of students who did not receive a Pell Grant at the institution and who completed in 2 years at original
LOAN_COMP_ORIG_YR2	Percent of students who received a federal loan at the institution and who completed in 2 years
NOLOAN_COMP_ORIG_YR2	Percent of students who never received a federal loan at the institution and who were still enrolled at original institution within 2 years
MD_INC_RPY_5YR_RT	Five-year repayment rate by family income ($30k-75k)
GRAD_DEBT_MDN	The median debt for students who have completed
WDRAW_DEBT_MDN	The median debt for students who have not completed
LO_INC_DEBT_MDN	The median debt for students with family income between $0 and 30k
MD_INC_DEBT_MDN	The median debt for students with family income between $30k and 75k
HI_INC_DEBT_MDN	The median debt for students with family income between over 75k
IND_DEBT_MDN	The median debt for independent students
PCTPELL	Percentage of Pell Grant
AVGFACSAL	Average faculty salary
TUITIONFEE_PROG	TUITIONFEE_PROG
faminc	Average family income
md_faminc	Median family income
mn_earn_wne_p10	Mean earnings of students working and not enrolled 10 years after entry
md_earn_wne_p10	Median earnings of students working and not enrolled 10 years after entry
Ethnies	
PBI	Flag for predominantly black institution
AANAPII	Flag for Asian American Native American Pacific Islander-serving institution
MENONLY	Flag for men-only college
WOMENONLY	Flag for women-only college
Selectivity	
ADM_RATE_ALL	Admission rate for all campuses rolled up to the 6-digit OPE ID
SATVR25	25th percentile of SAT scores at the institution (critical reading)
SATVRMID	Midpoint of SAT scores at the institution (critical reading)
SATVR75	75th percentile of SAT scores at the institution (critical reading)
SATMT25	25th percentile of SAT scores at the institution (math)
SATMTMID	Midpoint of SAT scores at the institution (math)
SATMT75	75th percentile of SAT scores at the institution (math)
SATWR25	25th percentile of SAT scores at the institution (writing)
SATWRMID	Midpoint of SAT scores at the institution (writing)
SATWR75	75th percentile of SAT scores at the institution (writing)
SAT_AVG_ALL	Average SAT equivalent score of students admitted for all campuses rolled up to the 6-digit OPE ID
ACTCM25	25th percentile of the ACT cumulative score
ACTCMMID	Midpoint of the ACT cumulative score
ACTCM75	75th percentile of the ACT cumulative score
University specs	
st_fips	FIPS code for state
region	Region (IPEDS)
locale2	Degree of urbanization of institution
OPEID	8-digit OPE ID for institution
CCBASIC	Carnegie Classification -- basic
CCUGPROF	Carnegie Classification -- undergraduate profile
UGDS	Enrollment of undergraduate degree-seeking students
PFTFAC	Faculty Rate

Figure 3: Codebook for our variables