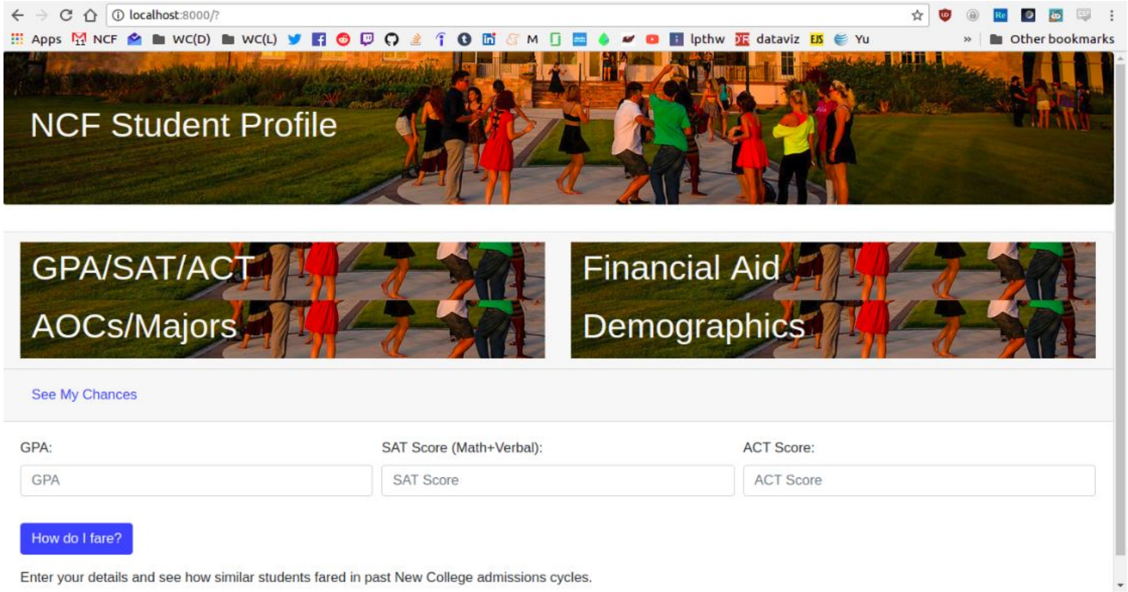
Design Document

Adriana Souza, Roger Filmyer, Lydia LaSeur

April 14, 2018

Contents



NCF Student Profile

GPA/SAT/ACT

AOCs/Majors

Financial Aid

Demographics

See My Chances

GPA: GPA

SAT Score (Math+Verbal): SAT Score

ACT Score: ACT Score

How do I fare?

Enter your details and see how similar students fared in past New College admissions cycles.

Figure 1: Current barebones of the visualization

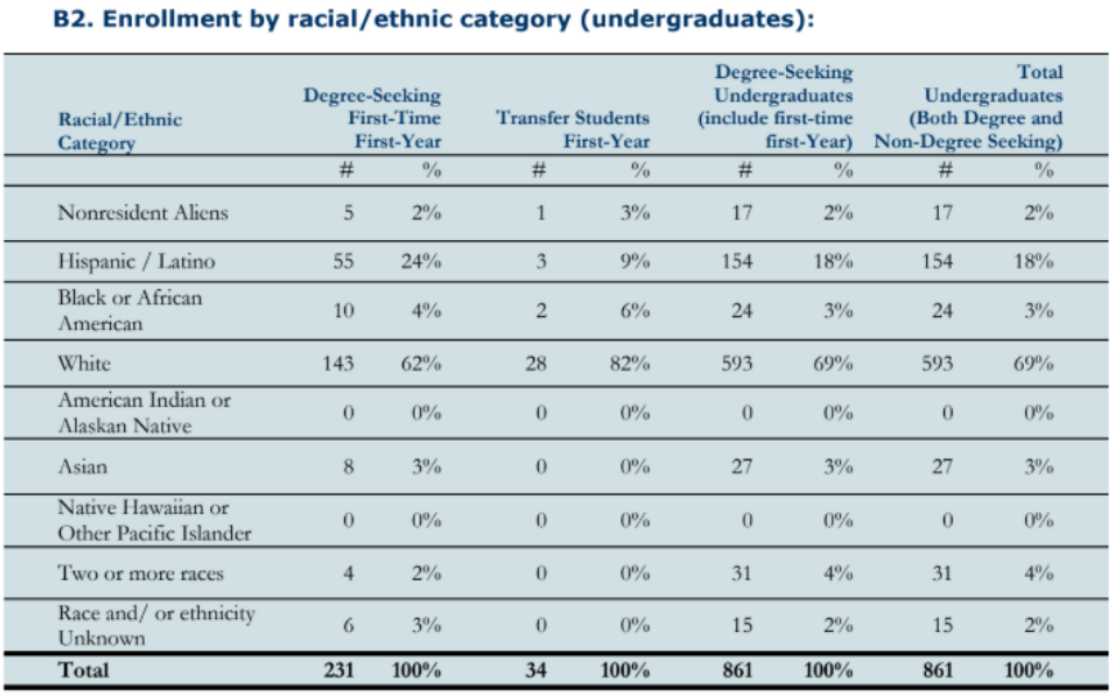
Introduction and background

College application time is a daunting period for everyone involved. For students, after 4 years of trying their best to be competitive, choosing the right college means sifting through tons of information that is either not collected in the same place or behind a paywall. For schools, it means making sure that all the information students need is available, up to date, and does not overwhelm.

This is the problem we are trying to solve; we created an interactive visualization that answers the most commonly asked questions by the students that uses data that may not be readily available on a common Admissions' page.

Every year, colleges prepare a Factbook – a reference document providing extensive descriptive data on the school it represents. This document holds the answer to what most students want to know during the time of the application: "What is the average student that gets accepted at this school like?", "Do I measure up?", "How many of those students get the type of financial aid I am looking for?", "Exactly how diverse is the school?", among others.

The data comes in a .pdf with a lot of tables that look something like this:



B2. Enrollment by racial/ethnic category (undergraduates):

Racial/Ethnic Category	Degree-Seeking First-Time First-Year		Transfer Students First-Year		Degree-Seeking Undergraduates (include first-time first-Year)		Total Undergraduates (Both Degree and Non-Degree Seeking)	
	#	%	#	%	#	%	#	%
Nonresident Aliens	5	2%	1	3%	17	2%	17	2%
Hispanic / Latino	55	24%	3	9%	154	18%	154	18%
Black or African American	10	4%	2	6%	24	3%	24	3%
White	143	62%	28	82%	593	69%	593	69%
American Indian or Alaskan Native	0	0%	0	0%	0	0%	0	0%
Asian	8	3%	0	0%	27	3%	27	3%
Native Hawaiian or Other Pacific Islander	0	0%	0	0%	0	0%	0	0%
Two or more races	4	2%	0	0%	31	4%	31	4%
Race and/ or ethnicity Unknown	6	3%	0	0%	15	2%	15	2%
Total	**231**	**100%**	**34**	**100%**	**861**	**100%**	**861**	**100%**

Figure 2: New College Factbook 2016-2017: Table B2

The answer to those questions are in these descriptive statistics that do not necessarily need inference, as much as a way to be represented such that they do not overwhelm. Compare the same information displayed in the graph below:

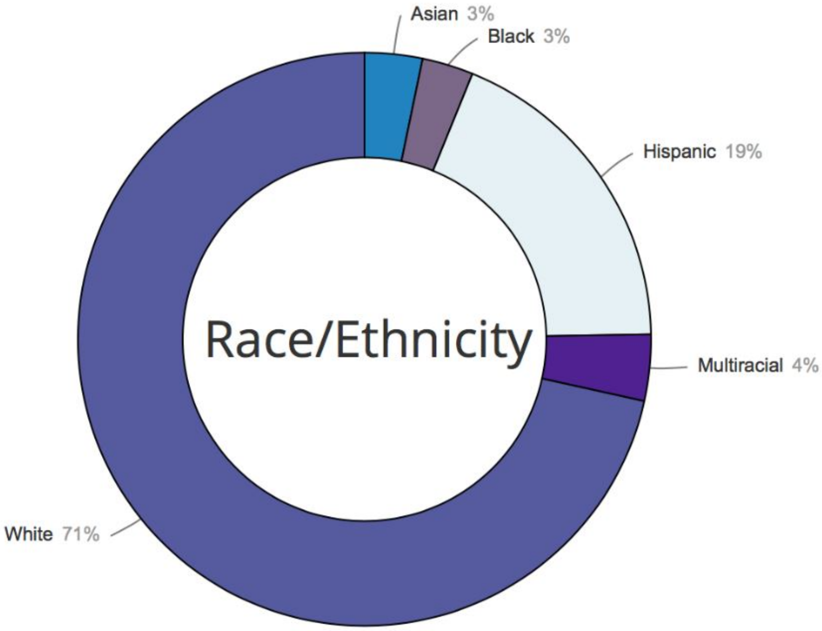


Figure 3: Visualization: Demographics Tab

Note: the colors are from a random palette, we will adjust to the website colors when embedding. The donut is also fully interactive. Given our audience and for simplicity, we excluded all races for which there were no students and also the “Unknown category.”

Data and outcomes

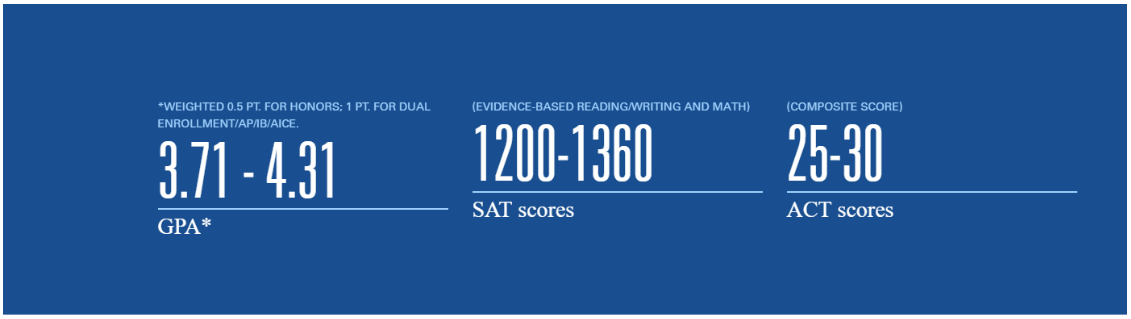
Structure of the data

There are two sources of data for two separate parts of the project. The two parts are (1) interactive visualizations with data on the average accepted student’s GPA, SAT, ACT, choice of major, financial aid received, and demographics; (2) an interactive tool where the prospective student can enter their GPA, SAT, or ACT and a reduced logistic model (written with stock javascript) will give them an idea how what their chances would be based on the same statistics for students who were accepted.

For the visualizations, we use the aggregate statistics reported in the Factbook, which is publicly available. For the second part, we used internal student data to calculate the coefficients (locally) that we passed to our model. This data is sensitive and contains student name, GPA, essay, and test scores.

Audience

The audience are prospective New College students and New College as well. Our goal is to improve on the current Class Profile of Students page (found here), by including the interactive visualization and making the part below interactive:



*WEIGHTED 0.5 PT. FOR HONORS; 1 PT. FOR DUAL ENROLLMENT/AP/IB/AICE.

3.71 - 4.31

GPA*

(EVIDENCE-BASED READING/WRITING AND MATH)

1200-1360

SAT scores

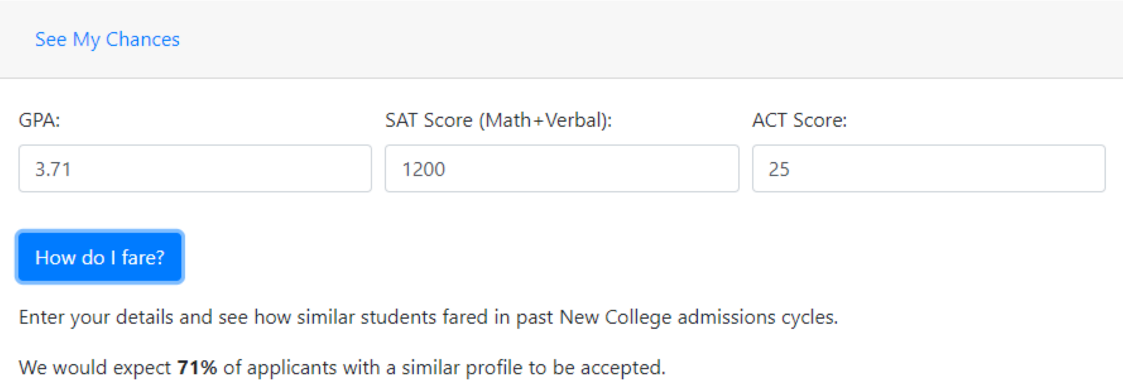
(COMPOSITE SCORE)

25-30

ACT scores

Figure 4: New College Class Profile Page

Right now, it looks like this:



See My Chances

GPA:	SAT Score (Math+Verbal):	ACT Score:
3.71	1200	25

How do I fare?

Enter your details and see how similar students fared in past New College admissions cycles.

We would expect **71%** of applicants with a similar profile to be accepted.

Figure 5: Visualization: "See my chances" section

After styling, it should look similar to the first picture, only interactive. The default values would be the ones reported, but the prospective student would be able to enter their own values and get an idea about the likelihood of getting accepted based on scores alone. There is a concern about how binding this information could be taken to be, so we will be sure to include instructions on how the tool is meant to be used and about the rest of the things considered during their application.

In essence, this presents the data in a clearer, individualized way. These numbers give you a vague sense of whether or not you, as a student, would be accepted but it offers no degree of differentiation from other schools' website. There is value in interactivity. The prospective student has a better experience when they are able see their number, rather than an impersonal range. The student gains and so does the school.

Features

Our visualization collapses and expands according to what information the user wants at the time. In the end, the user is able to print their page (to a .pdf or actual prints) and keep only the data they are most interested in.

Design process

Pros and cons

The NCF Factbook provides data pertaining to the high school performance and demographics of enrolled students, academics and faculty at NCF, cost of attendance, and financial aid. All of this is presented using tables. These tables are useful for pulling exact values but they are tedious to read and don't communicate the overall distributions in a quick and easy manner. Since our overall goal is to provide prospective freshman with a quick and easily understood summary of statistics we will translate these tables into graphs that communicate the overall distributions of these variables in just a glance.

Our data consists of three different types: numerical, categorical, and geographic. For each data type we will consistently use a specific type of graph. We chose to keep our visualizations simple and clean to ensure that they are readable by a broad audience.

1. **Numerical - Histograms/Bar Charts**

   Variables that fit this type include GPA, ACT and SAT scores, high school class rankings, and class sizes. For each of these variables the values are binned and each bin is assigned either a count or a percentage. The factbook has tables for GPA, ACT/SAT scores but we could not use them for reasons that will be discussed in the next section. As a result we used the internal dataset provided by Admissions to create aggregate counts for binned values to use in histograms.

   Pros:

   - The ordering of the bins will be preserved.
   - Only aggregate statistics are needed, allowing us to use certain variables from the internal dataset without but still maintain anonymity.
   - The format is very familiar and easy to read for a broad audience. Simple enough for us to automate our D3 code to work for several different variables.

   Cons:

   - While they're effective, they're not very exciting or interactive for the user.
   - Variables with very sharp distributions where the majority of the population lies within a few bins don't translate well.

2. **Categorical - Donut/Pie Charts**

   Most of the Factbook data is categorical, including demographics for both students and faculty as well the academic offerings at NCF.

   Pros:

   - There is no ordinality to this data so we decided to not use histograms to avoid any implication of hierarchy from the ordering of the categories along the axis.
   - They fulfill our goal of providing a quick, overall view of the distribution.

- Just like histrograms, they are very familiar and any user should already know how to interpret it.
- Simple enough to automate our D3 code which is very important since a majority of our variable are categorical.

Cons:

- It's hard to accurately judge the differences in area between each segment or slice.
- User can't read any exact values from it unless a legend or individual labels are provided, which can get messy for variables with a lot of possible categories or categories with long names.

3. **Geographic - Choropleth Maps**

   We have two geographic variables: Florida county of residence for in-state students and US region of residence (e.g. New England, Midwest, South excluding Florida, etc). For these we will use choropleth map to illustrate the distribution of students across these locations.

   Pros:

   - Obviously a map is the easiest way to communicate any geographic data.
   - Most may not know choropleths by name but they're still familiar and easy to read for the general public.
   - For our purposes, there are pre-existing appropriately drawn maps so the implementation is not difficult.

   Cons:

   - The individual regions in county map of Florida will be too small to insert any names or values and some interactivity will be required to provide that information in a clean visualization.

Below is an outline of the current architecture of the visualization. Each numbered group will have its own dedicated section to hold the individual graphs of its variables. For each variable, the type of data, and possible values/categories.

1. GPA and Test Scores
   - GPA - Numerical - Histogram
     Possible Values: Half point bins on a 5 point scale.
   - HS Class Rankings - Numerical - Histogram
     Possible Values: Top 10th, 25th, 50th percentile or bottom 50th or 25th percentile.
   - SAT Scores - Numerical - Histogram
     Possible Values: 1000 point bins on a 1600 point scale.
   - ACT Scores - Numerical - Histogram
     Possible Values: 3 point bins on a 36 point scale.
2. Student Demographics
   - Gender - Categorical - Donut Chart
     Possible Categories: Male or Female
   - Race/Ethnicity - Categorical - Donut Chart
     Possible Categories: White, Black, Hispanic, Asian, Multiracial
   - Residency by US Region for all students - Geographic - Choropleth
     Possible Locations: New England, Midwest, Middle States, West, South (minus FL), Southwest,

Florida

- Residency by County for in state students - Geographic - Choropleth
  Possible Locations: All counties in FL

- Secondary School Type - Categorical - Donut Chart
  Possible Categories: Public, Charter, Private, Parochial, Home, GED, International

- Specialized HS Curriculum - Categorical - Donut Chart
  Possible Categories: Honors, AP, IB

3. Academics
   - Distribution of AOC's by Academic Division (Graduates) - Categorical - Donut Chart
     Possible Categories:- Social Sciences, Humanities, Natural Sciences, Interdisciplinary Studies, Environmental Studies, General Studies

   - Distribution among Top 6 AOC's (Enrolled Students) - Categorical - Donut Chart
     Possible Categories: Biology, Chemistry, Psychology, Political Science, Economics, Humanities

   - Class Sizes and Student to Faculty Ratio - Numerical - Histogram
     Possible Values: 2-9, 10-19, 20-29, 30-39, 40-49, 50-99, or 100+ students enrolled
     Student to faculty ratio will be included with the histogram for class sizes since the two are closely related and it seems unnecessary to provide a separate visualization for a single ratio.

   - Faculty Demographics - Categorical - Donut Chart
     Possible Categories: Gender and Ethnicity are provided by the factbook, using the same possible categories to describe students.

4. Financial Aid and Cost of Attendance
   We are still waiting to see if we will get the financial aid data so we can't give exact details. If possible, we would like to provide statistics on the types of aid provided and the number or proportion of students that received each type. We also plan to show the estimated cost of attendance using either a bar chart or donut chart.

Implementation challenges

1. **Display: none**

   We had several aesthetic issues with d3 graphs. Since our whole visualization depends on being collapsable, a lot of the elements were being rendered while invisible to the user. This resulted in text not wrapping correctly and made it difficult to create graphs with sizes that would adjust automatically to the page. Our solution was to make our graphs non-resizable and by adding some esoteric javascript event handling magic from the sacred texts (Stack Overflow) to check for changes in the webpage and dynamically change the width and height of the graphs.

2. **How to get all SATs under the same scoring band after the changes in 2016?**

   Changes were made to the SAT in March 2016 (details of which will be discussed later) and the factbook data uses the old scoring metrics. Any user using this product will have taken the newer version and they can't accurately compare themselves to NCF students if we provide data for the older version. Plus, the internal dataset used to build our model included students that had taken both the pre March 2016 version, the post March 2016 versions and some applicants had both. To make sure our data was both relevant and consistent we used a conversion chart provided by College Board to convert all of the old SAT scores to their equivalent new score.

3. **How to reconcile the inconsistent GPA scales between the Factbook data and internal data?**

   The Factbook has GPA's on a 4 point scale but NCF uses a weighted GPA on a 5 point scale when evaluating an applicant and consequently our regression model does the same. We want to maintain consistency throughout our product and felt that mixing the two different GPA scales would be confusing for the user. NCF provides instruction on how they weight high-school GPA's so any prospective student could recalculate their own if needed. Using the internal dataset we generate our own binned counts so our graphs would show GPA's on the same scale used in our model.

4. **Why are my coefficients wrong?**

   When we first ran the regression, our coefficient for GPA seemed to give it a lot less weight than the guidelines for acceptance decision dictated (in theory). Upon further inspection, we realized IB students, who do not have a regular GPA, were being flagged internally as having a 9.8 GPA. This completely threw off our logistic model. When we set those to NAs, the essentially coefficient doubled and our results were significantly more in line with what the data showed.

5. **How to automate D3?**

   Our plan includes around 10 to 12 D3 graphs. Most d3 examples only deal with a single, isolated graph. We had to look into a way of being able to create these graphs programmatically. How do we productize graphing? How do we avoid DRY issues? We wrote a function that handles all of that beautifully if you just pass it a .csv file with the data you want to display in your bar chart.

6. **How to word the results of our model in an informative way but that is also clearly non-binding?**

   We did not want to report the prediction from our model in a way that would imply that the prediction was guaranteed or that we were providing the true probability of acceptance. We chose to phrase the results as the percentage of admitted students that had similar or worse scores than the input.

7. **Why make the graphs scale well with the webpage's responsive design (also be mobile friendly, etc)?**

   We envision a large number of our prospective students and parents using our product on their smartphones. How can we deliver a mobile-friendly experience? This was one of our main challenges that was taken into consideration as we moved onto the final stages of the .CSS styling process.

Incorporating feedback

After our presentation, the class expressed that if they were to use such a visualization, there were 3 other things they would like to see incorporated that we did not mention: (1) number of students who are the first generation in college; (2) to include more specific financial aid information such as number of Pell grants, Stafford loans, Bright Futures, and Fullbrights; and (3), number of students placed in graduate programs.

We have reached out to the financial aid department regarding the number of students admitted last year who were the first generation in their family to be in college. We are still currently waiting for that. For point (2), we scraped and included information about all of the types of financial aid except Fullbrights; we were not able to find a good source for this number. Lastly, for point (3), this information is very hard to track down for many reasons. The school does have a survey that is administered at graduation (regarding future plans of enrollment in graduate school) as well as one that is sent out a year after (to check up on that information), but the response rates are low. That, coupled with the fact that most students move out of state or take a gap year, makes it even harder to be able to narrow down that information.

Evolution of the design process

Both our goals and audience changed as we refined the design process. Initially, we had intended to design a model for the admissions department and make some visualization of how the different components of an application affect the acceptance decision. As we studied New College's decision method, this changed: we started thinking about how each school is different when it comes to what they consider important. What kind of information do students need when applying? Do they know what it is? If so, where do they get it?

The answer to those questions usually lies behind a paywall. There is an entire industry focused around this exact problem. We could not find any free alternatives that do this, so we decided to create our own tool for prospective New College students.

Conclusion

Goals, functionality, limitations

Our visualization condenses key pieces of information about New College for prospective students and parents. The user will be able to have the information they need in one place, without having to go through tables and tables of irrelevant information contained in the factbook. The major goal is to inform, in a way that does not overwhelm. This is especially important since most students are not aware of the existence of Factbooks. These are freely distributed by colleges, with almost the exact same information submitted to places like Barron's and Princeton Review, etc – whose services the students often pay for.

Before the final submission, we expect to have the styling completely done. As of now, we have only taken care of the the barebones and base functionality. The design aspect is currently in the works. In the future, the tool would benefit from having a more complex model where we would be able to incorporate course rigor based on what school the student selects, for example.

Division of work:

Roger handled all of the front end aspect of the project. He also wrote the logistic regression using stock js and the function we used to create bar charts. Lydia collected and cleaned the data, she also corrected the SAT scores so that they would be on the same scoring band. She also created some of the graphs on the page. Adriana was the project manager and pitched in a little bit everywhere.