COLLEGE SEEKER: VISUALIZING COLLEGE SIMILARITY AND PREDICTING STUDENT DEBT

Austin Krauss, Sanjana Kumar, Stephen Mullaly, Dan Schauder, Matt Schlosser

INTRODUCTION

While college choice is a critical life decision, there is a lack of resources that help prospective students identify potential college options as well as consider their associated long-term financial impacts. Our team has addressed this gap with a novel interactive web app (henceforth referred to as collegeseeker.net) to explore college choice and associated debt repayment. Specifically, collegeseeker.net finds and filters colleges based on similarity then details a regression model showcasing students' ability to avoid long term student debt issues.

Existing resources include the US Dept. of Education College Scorecard, publications like the US News and World Report, and commercial websites like collegesimply.com. Recent research found that the usefulness of some of these sources is limited [1], and none of these solutions provide an intuitive or insightful visual experience that is critical for attracting and engaging teenage users.

Student debt has created a rapidly-escalating crisis in the United States. Students often ignore the harsh realities of student loans and debt repayment, but focus on qualitative factors like university prestige and extracurriculars [2]. These factors, particularly prestige, may not benefit their future career and earnings potential but can result in higher debts [3]. In addition, research shows that 78% of students underestimate the total cost of their loans [4]. The result, as of 2019, is 43 million Americans with federal student loans and almost 20% in default [5]. Early exposure to financial literacy resources can lead to a significant reduction in risk of student debt default [6], and our product aims to fill this need.

PROBLEM DEFINITION

Ideally, students apply to a number of schools and can be confident in pursuing their college goals without the concern of debt. In reality, students often overlook schools similar to their dream school and succumb to crippling loan debt upon graduation. The consequences of student loans impact the lives of those repaying them by contributing to financial challenges and hindering future educational pursuits. Our website, collegeseeker.net, will allow students to understand their college choices and gauge students' general ability to repay student loans which in turn will allow these prospective students to make more educated decisions when selecting a four-year university.

SURVEY

Several sources concerning college choice and debt were reviewed to understand current and long standing implications. Research shows that a college graduate's backgrounds and experiences affect their student debt burden ratio by revealing certain factors that have a dramatic impact on debt, such as choice of major, degree of urbanization, and whether the university was public or private [7]. Furthermore, the risks of carrying debt are significantly differentiated by gender as women stay in school longer and drop out at lower rates than men, thus incurring a larger burden of debt [8]. Researchers also found a link between increased debt and a delay in starting a family [9] as well as being more likely to have lower wage growth [10]. Finally, debt from a public college significantly reduces the odds that somebody will pursue an MBA, doctoral degree, or first professional degree [11].

We further considered existing research on similarity and regression. For our continuous data sets, Minkowski Distance of order 2 (Euclidean) is the most widely used distance measure [12]. Specifically, using weighted Minkowski distance can help detect ordinal association with proper weighting [13].

METHOD

Data

In order to help students explore college options and understand the possible debt associated with various schools it was important to have consistently collected data for schools across the United States. Fortunately, the U.S. Department of Education curates a "College Scorecard" database with the information needed for developing these tools. The full scorecard database covers 23 years of datasets and is 4.89 GB. It is freely available for download from the Department of Education website. For this project, the focus was on the year 2019 as this is the most recent dataset available. When viewed on its own, the data for 2019 is a 168 MB CSV file with 2,384 columns and 6,806 rows for a total of 16,225,504 cells. Data was appropriately filtered and values were imputed where necessary using generally accepted techniques appropriate for the column's data type [14].

After filtering, the data was used to create word clouds to illustrate the popularity of majors offered at each school. Columns in the dataset provided the proportion of degrees awarded in each field of study for each school. Using these columns, a dictionary was created for each school which contained the percentage of degrees awarded in each program, rounded to the nearest integer. A word cloud was generated using the values of the dictionary to size the majors. All of the word clouds were generated and uploaded to a Google Cloud Platform bucket to be referenced by the College Seeker tool.

The data also contained a column with each university's homepage URL. This was used along with the Clearbit.com/logo API to return a logo for each school. The API used is intended to be embedded in the HTML of a website and call logos on demand, but due to the number of logos being used a script was written that would call each university logo and upload it to a Google Cloud Platform bucket. This made every logo available for immediate use. Missing logos were replaced with a generic logo and a list of schools using this logo was output to a text file. Of the 1,735 logos needed, less than 100 were missing and they were manually included.

Discovery Graph (Innovation #1)

To aid students in discovering alternative educational institutions, we built a similarity network in which each university is a *node*, and similar universities have *edges* that connect them with the strength of an edge determined by the similarity of the two schools. This innovation enables the student to discover new universities that they may have otherwise missed in their school search. Additionally, the user has the flexibility to add more weight to certain characteristics of the schools. For example, if they prefer highly selective schools in the northeast region of the United States, they can increase the weights of those dimensions. The graph is created by defining a vector space for each university, which includes variables like geographic location, size, degrees offered, and average SAT score. Then, we created a pairwise distance matrix using weighted Minkowski distance with a p-norm of 2. This is similar to the Euclidean distance with the added complexity of using the weights to differentiate the importance of each variable (see Appendix for mathematical formula).

Regression (Innovation #2)

The second key component of our analysis was to enrich the discovery graph with predicted loan repayment rates. To calculate these rates, we trained a multivariate regression model using institutional features to predict students' ability to repay educational loans. First, we divided our dataset into training and testing datasets. The training dataset was used to train and cross-validate our regression model, while the testing dataset was used for the predictive task. To improve our model efficiency, we used "Grid Search Cross Validation and Random Search Cross Validation" for hyper parameter tuning of the linear regression model [15]. Using the best hyper parameters, we evaluated our model's performance using cross validation and RMSE score. Finally, we used this model to predict the 5-year loan repayment rate for the testing dataset. With this innovation, collegeseeker.net enables the student to select schools by gauging past students' ability to repay their loans at that particular school.

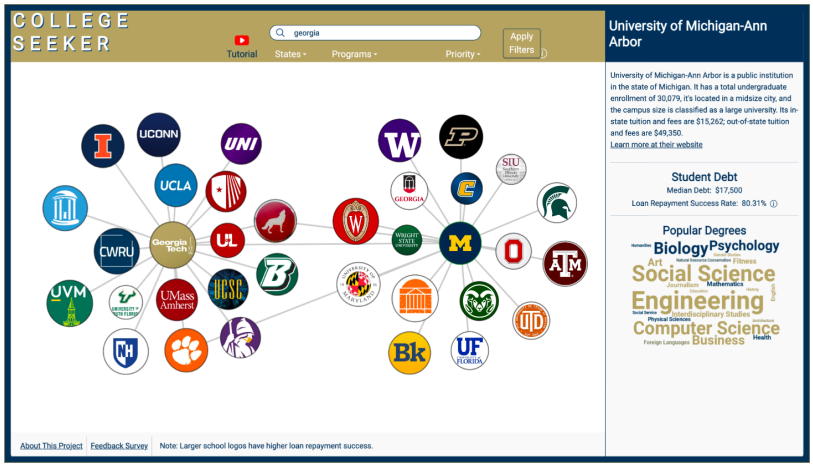
Visualization (Innovation #3)

The discovery graph and the regression analysis are combined to create an interactive visualization. Students may begin their exploration by providing an initial school of interest or by beginning at a default location. A set of dropdown filters provide the user with the flexibility to filter the pool of schools in the visualization by state and program of study. Additionally, students have the option of adjusting the weights used in calculating the pairwise distances that form the graph, allowing them to emphasize selectivity, geographic location, and campus experience.

As there are well over 1,500 universities included in the discovery graph, we constrain the view to show roughly 15 nearest neighbors of a given university at a time. Upon double-clicking a neighboring university, its neighborhood is expanded, and the view pans and zooms to focus on the new neighbors that were revealed. This low-level exploration of individual university "clusters" allows for a more immersive experience and avoids the infamous dilemmas associated with "hairball" style network visualizations. Additionally, each node in the visualization is sized by the loan repayment rate predicted by our regression analysis.

Lastly, with a single click on a university node, school specific information, including debt and a word cloud appears. The programs of study at the university appear in the cloud with each program being sized by the proportion of undergraduate degrees awarded at the institution.

We believe that the visual experience described here allows students an intuitive way to compare a pool of universities they may want to consider on the basis of successful loan repayment at a glance. We hope this approach proves insightful for students by introducing them to universities they may have otherwise overlooked while tempering their expectations with a clear depiction of the ability to repay student loans.



EXPERIMENT AND EVALUATION

We conducted a randomized experiment to evaluate our approach. 104 college-bound juniors from McKinney Boyd High School in McKinney, Texas (suburb of Dallas, Texas) were randomly assigned to interact with either the College Seeker website or U.S. News's college search website (after informed consent was obtained from the parent). Each student received similar instructions and 10 minutes to use the website. After the set time was concluded, students were given a brief survey. By controlling confounding variables and randomly assigning subjects to treatment groups, we worked to minimize bias in our data and allow for causation to be established.

The experiment measured four response variables: college discovery, awareness of loan repayment rate, awareness of popular majors and programs, and willingness to recommend the website to peers. In order to reduce variability in our response variables, the experiment controlled potentially confounding variables. To this end, the following text was given to each treatment group:

[collegeseeker.net / usnews.com/best-colleges/college-search]

Thank you for volunteering to participate in this brief experiment! The researchers are depending on your serious consideration, so please remain focused on this task and nothing else for the next fifteen minutes. Follow these exact instructions and do NOT click ahead before Schloss Boss says so. Please do not speak to any of your peers during this experiment.

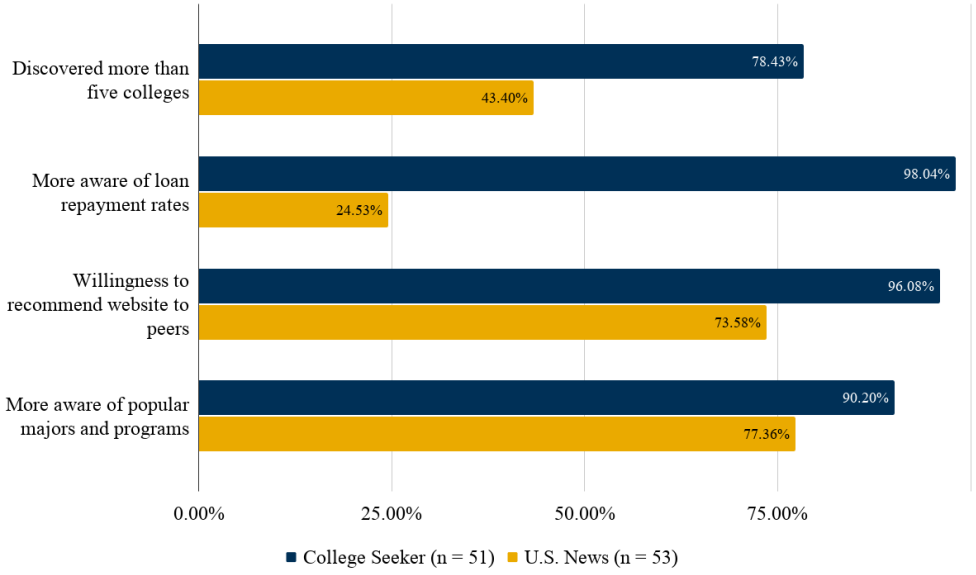
Go to [collegeseeker.net / usnews.com/best-colleges/college-search]. Interact with this website and only this website until time expires. Note: If you click on a school's website, please return to the original website ASAP to continue.

[After 10 minutes expires] Please candidly answer the following questions:

1. How many colleges did you discover using the [College Seeker or U.S. News] website?
2. After using the [College Seeker or U.S. News] website, are you more aware of the percent of students who repay their loans?
3. After using the [College Seeker or U.S. News] website, are you more aware of the most popular majors at the colleges you investigated?
4. Would you recommend the [College Seeker or U.S. News] website to your friends when researching colleges/universities?

Here are the results from our experiment:

Figure 3: Results of Experiment	Discovered more than five colleges	More aware of loan repayment rates	More aware of popular majors at particular colleges	Recommend website to research colleges
College Seeker (n = 51)	78.43%	98.04%	90.20%	96.08%
U.S. News (n = 53)	43.40%	24.53%	77.36%	73.58%
95% Confidence Interval for the difference in College Seeker and U.S. News	(5.322, 11.132)	(0.613, 0.857)	(-0.011, 0.268)	(0.095, 0.355)



Other considered methods were a matched pairs experimental design in which each subject would serve as their own control. The major drawback would be the influence of the treatment order. If a subject used College Seeker first then U.S. News second, this might skew the results in favor of the most recent treatment (recency bias). Similarly, if a subject first engages U.S. News then College Seeker, they might

naturally prefer College Seeker. Lastly, by using a large number of subjects we can assume that confounding variables are evenly spread amongst the two groups. This allows for an effective comparison of results without the inherent bias of a matched pairs experimental design.

Survey Feedback

Valuable feedback was received from a survey of 40 college-bound high school juniors from McKinney Boyd High School in McKinney, Texas that used collegeseeker.net. The feedback was mostly positive but we were able to learn from some of the negative feedback as well. Most of the negative feedback related to site layout on lower resolution monitors which was investigated and rectified wherever possible.

A couple comments made it seem that more focus on motivation may be beneficial for students. Some students may not understand the concerns of student debt or think it doesn't sound like a problem until they are faced with it upon graduation. While we don't know the situation of the anonymous survey takers the following results may reflect this:

"Debt isn't one of my major concerns when deciding on a college at the moment."

"I'm not really interested in the percentage of people who repaid their loans"

When thinking through how to share motivation for why debt is an important consideration it may also be necessary to do work to make the site even more user friendly:

"It is a great idea to make information about different colleges easily accessible to highschool students. The only critique that I have is that I feel that a website should strive to be simple and direct. In my opinion, the instructions on how to use the website should be in the website.(maybe short tips next to important buttons) I wouldn't want to have to watch a video on how to use a website, especially when it seems that the website is meant to be easy to use."

The majority of positive feedback aligned with the goals of the site. This showed that many students appreciated the ability to see information on many schools all in one place and it helped them to expand the number of schools they were considering and to include debt in the decision making process:

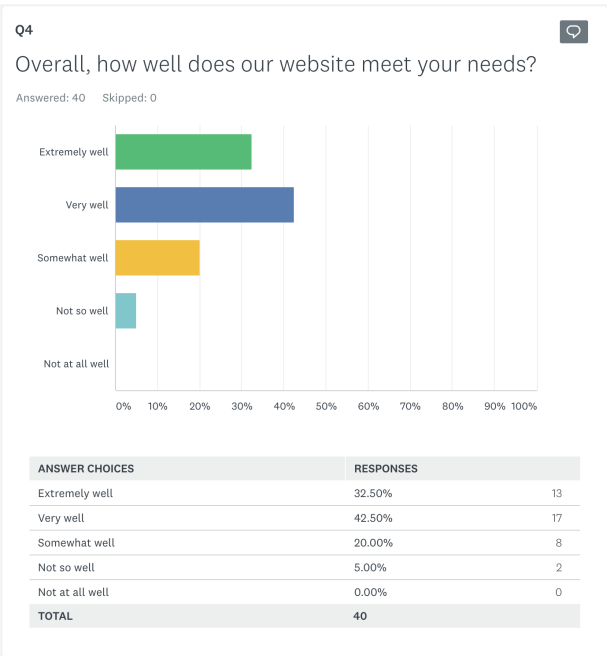
"I liked everything about the website, every part of the page is easy to use and understand, which is extremely helpful when information about colleges can be overwhelming."

"Helped me see what majors/degrees are most popular at each college."

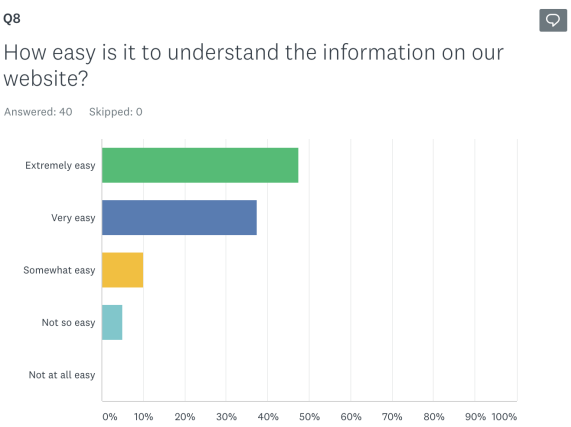
"It's really difficult to go from college to college looking at their admissions information, so it's really helpful looking at it all in one place."

"Seeing the debt helps you understand if people are able to get jobs at the end and pay off their loans, helps when deciding the money aspects of college."

"Having the student debt statistics be so clear and accessible is very helpful in the college picking process, it is also very handy to pick where you would like to be located and what you would like to study in order to narrow down your choices. This website makes deciding what colleges to put at the top of my wishlist a lot easier!"



ANSWER CHOICES	RESPONSES	
Extremely well	32.50%	13
Very well	42.50%	17
Somewhat well	20.00%	8
Not so well	5.00%	2
Not at all well	0.00%	0
TOTAL	40	



ANSWER CHOICES	RESPONSES	
Extremely easy	47.50%	19
Very easy	37.50%	15
Somewhat easy	10.00%	4
Not so easy	5.00%	2
Not at all easy	0.00%	0
TOTAL	40	

LIMITATIONS

Please note that our team is composed of working professionals and full-time students, so we have had to prioritize certain features and developments to match our available resources and time constraints. As such, the website has not been optimized for mobile use and may be slow to load. Additionally, due to the aforementioned limitations, we did not have time to optimize the back-end infrastructure to efficiently filter and render large amounts of data to the front-end in a timely manner. This means that we only store data on each university and its closest "neighbors", so the viewer may find that, if they select very specific filters (ex: limiting their search to only one state), only a few "neighbors" will show for a given university. If we had further bandwidth or time for development, these would be high on our priority list of future enhancements.

CONCLUSION AND DISCUSSION

In this project, we built an interactive website which allowed prospective students to understand their college choices and gauge students' general ability to repay student loans.

There are three main innovations for our project: First, this project innovatively comes up with the idea to design a college recommendation system that allows prospective students to make more educated decisions when selecting a four-year university. Secondly, we trained a multivariate regression model using institutional features to predict students' ability to repay educational loans. Finally, our user interface is interactive and user-oriented. Our project will definitely contribute tremendous value in helping prospective students identify potential college options as well as consider their associated long-term financial impacts.

TEAM CONTRIBUTIONS AND ROLES

All team members have contributed a similar amount toward completing the project.

Team Member	Austin	Sanjana	Stephen	Dan	Matt
Role	Data Engineering	Modeling/ Back End	Modeling/Back End	Visualization/ Front End	Visualization/ Documentation

REFERENCES

1. Mabel, Z., Libassi, C. J., & Hurwitz, M. (2020). The value of using early-career earnings data in the College Scorecard to guide college choices. Economics of Education Review, 75, 101958.
2. Espinoza, S., Bradshaw, G., & Hausman, C. (2002). The importance of college choice factors from the perspective of high school counselors. College and University, 77(4), 19.
3. Brand, J. E., & Halaby, C. N. (2006). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. Social Science Research, 35(3), 749-770.
4. Macy, A., & Terry, N. (2007). The determinants of student college debt. Southwestern Economic Review, 34, 15-25.
5. House, T., & Plus, G. I. Student Loan System Presents Repayment Challenges.
6. Fox, J. J., Bartholomae, S., Letkiewicz, J. C., & Montalto, C. P. (2017). College student debt and anticipated repayment difficulty. Journal of Student Financial Aid, 47(2), 111.
7. Chen, R., & Wiederspan, M. (2014). Understanding the determinants of debt burden among college graduates. The Journal of Higher Education, 85(4), 565-598.
8. Dwyer, E., Hodson, R., Mccloud, L. (2012) Gender, Debt and Dropping out of College.
9. Erin Velez, Melissa Cominole & Alexander Bentz (2019) Debt burden after college: the effect of student loan debt on graduates' employment, additional schooling, family formation, and home ownership, Education Economics, 27:2, 186-206.
10. Minicozzi, A. (2005). The short term effect of educational debt on job decisions. Economics of Education Review, 24(4), 417-430.
11. Zhang, L. (2013). Effects of college educational debt on graduate school attendance and early career and lifestyle choices. Education Economics, 21(2), 154-175.
12. Boriah, S., Chandola, V., & Kumar, V. (2008, April). Similarity measures for categorical data: A comparative evaluation. In Proceedings of the 2008 SIAM international conference on data mining (pp. 243-254). Society for Industrial and Applied Mathematics.
13. Zhang, Q. (2019). A Class of Association Measures for Categorical Variables Based on Weighted Minkowski Distance. Entropy, 21, 990.
14. Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, *4*(1).
15. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

APPENDIX

Pairwise distance matrix using weighted Minkowski distance with a p-norm of 2:

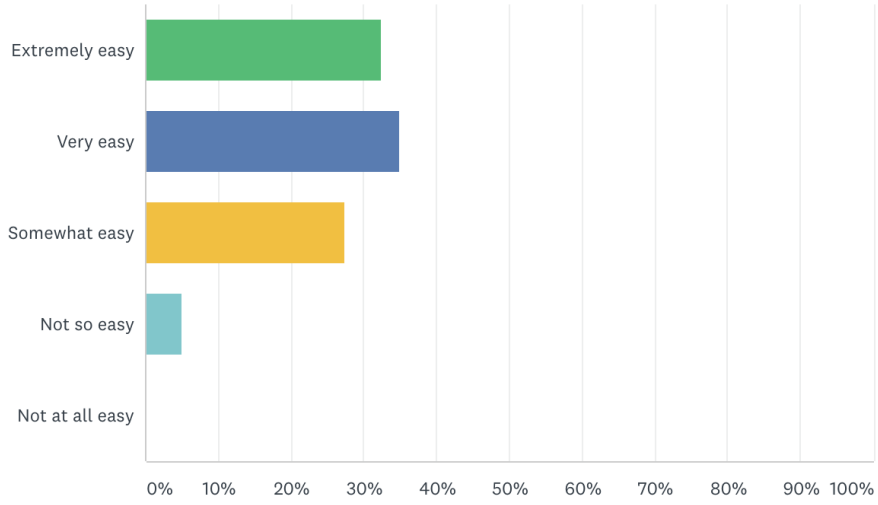
$$(\Sigma\,(|w_i(u_i - v_i)|^p))^{1/p}$$

Additional feedback from student survey follows:

Q5

How easy was it to find what you were looking for on our website?

Answered: 40 Skipped: 0

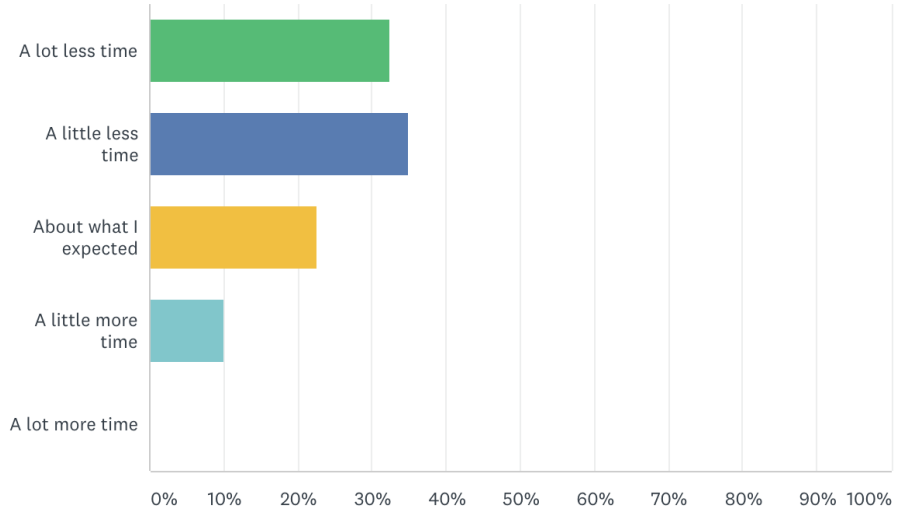


ANSWER CHOICES	RESPONSES	
Extremely easy	32.50%	13
Very easy	35.00%	14
Somewhat easy	27.50%	11
Not so easy	5.00%	2
Not at all easy	0.00%	0
TOTAL	**40**	

Q6

Did it take you more or less time than you expected to find what you were looking for on our website?

Answered: 40 Skipped: 0

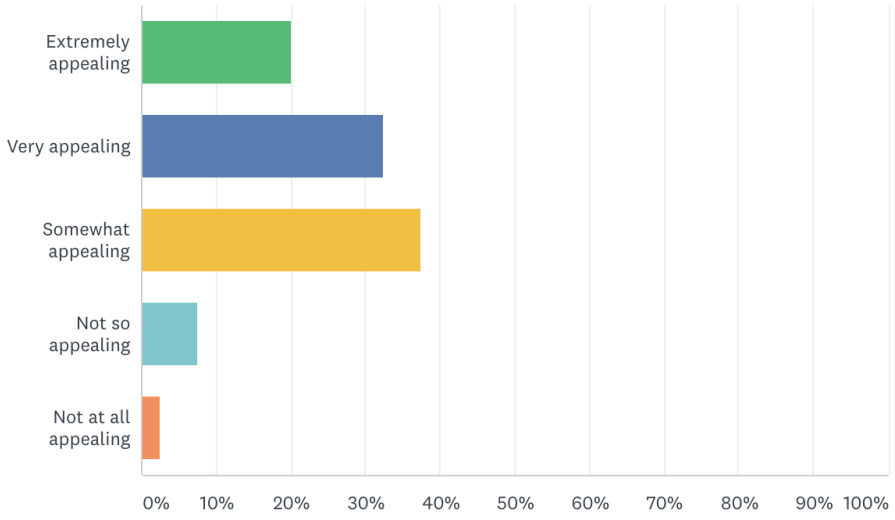


ANSWER CHOICES	RESPONSES	
A lot less time	32.50%	13
A little less time	35.00%	14
About what I expected	22.50%	9
A little more time	10.00%	4
A lot more time	0.00%	0
TOTAL	**40**	

Q7

How visually appealing is our website?

Answered: 40 Skipped: 0

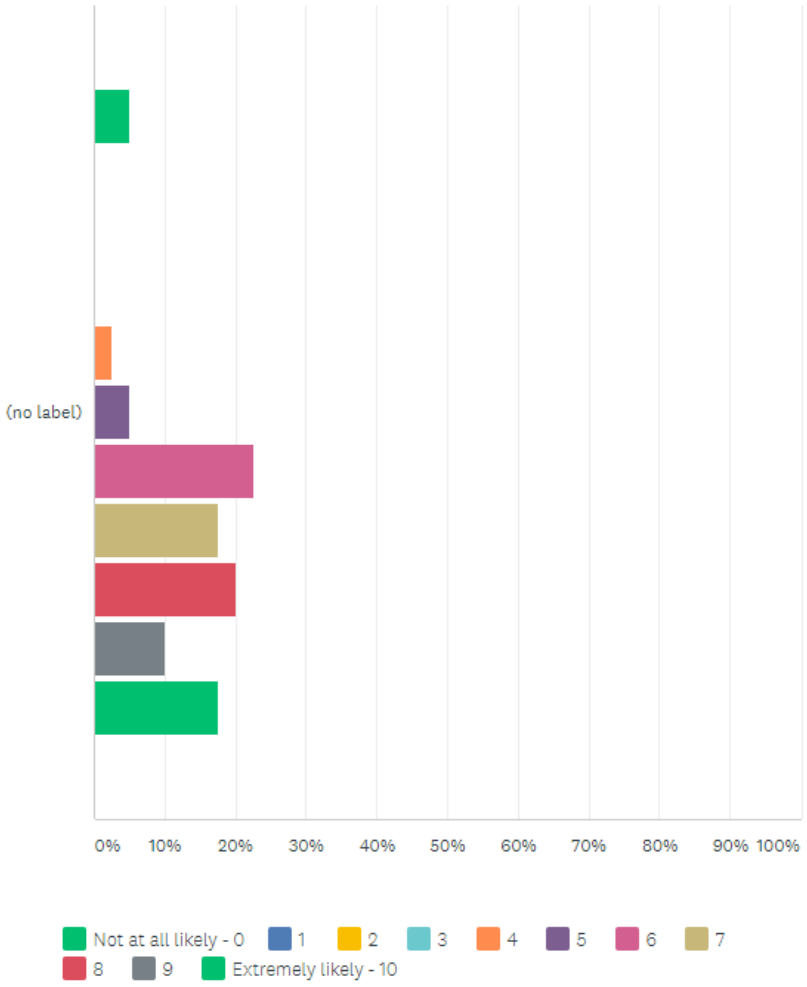


ANSWER CHOICES	RESPONSES	
Extremely appealing	20.00%	8
Very appealing	32.50%	13
Somewhat appealing	37.50%	15
Not so appealing	7.50%	3
Not at all appealing	2.50%	1
TOTAL	**40**	

Q9

How likely is it that you would recommend our website to a friend or colleague?

Answered: 40 Skipped: 0



	NOT AT ALL LIKELY - 0	1	2	3	4	5	6	7	8	9	EXTREMELY LIKELY - 10	TOTAL	WEIGHTED AVERAGE
(no label)	5.00% 2	0.00% 0	0.00% 0	0.00% 0	2.50% 1	5.00% 2	22.50% 9	17.50% 7	20.00% 8	10.00% 4	17.50% 7	40	-7.50