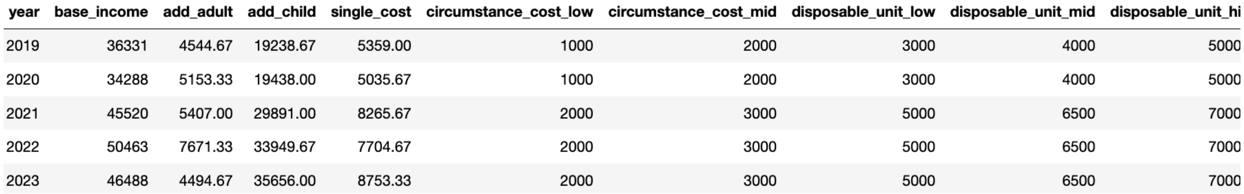
Final Report and Summary
Scholarship Allocation Tool for Brown Girl Surf
Rebecca Sandidge, PhD: Springboard Project 3

Problem:
Brown Girl Surf is a non-profit, community-based organization in Oakland, California that works to increase access to the ocean for women and non-binary people of color. BGS offers a summer youth leadership development program called “Rising Leaders”. Participants can apply for scholarships to cover up to 90 percent of the $900 tuition. Applicants submit basic information about their household income, family size, and they can include open format notes indicating additional financial needs faced by the household. Past funding allocation was determined manually, by employees who are pressed for time, based on conversations with parents and application data. The process was time-consuming and the subjective approach resulted in uneven distributions. The Scholarship Allocation Tool was developed as an algorithm that determines funding eligibility for each applicant using data they provide combined with data from the web. It applies mathematical computations to assess need and recommend a funding amount. It processes applications in very little time and allows the organization to report allocations quickly with confidence that funding distribution is fair and criteria are clear.

Methods:
Algorithm:
Need assessment for each applicant is based on living wage values for Alameda County, California. These data are scraped from an MIT website that updates required incomes for various family sizes and structures each year for every county in the United States. The tool scrapes past years and current values, calculates cost of additional family members, and stores these values for use by the funding algorithm (Table 1). A required family income value is calculated for each applicant based on applicant data. This value is used to determine how much disposable income the household has available. Disposable income is divided based on family structure and the amount is used to assign the applicant to one of five scholarship funding tiers: 90, 80, 70, 50, or 0 percent funding.



year	base_income	add_adult	add_child	single_cost	circumstance_cost_low	circumstance_cost_mid	disposable_unit_low	disposable_unit_mid	disposable_unit_hi
2019	36331	4544.67	19238.67	5359.00	1000	2000	3000	4000	5000
2020	34288	5153.33	19438.00	5035.67	1000	2000	3000	4000	5000
2021	45520	5407.00	29891.00	8265.67	2000	3000	5000	6500	7000
2022	50463	7671.33	33949.67	7704.67	2000	3000	5000	6500	7000
2023	46488	4494.67	35656.00	8753.33	2000	3000	5000	6500	7000

Table 1. Cost of living data; base_income is for one adult with no children.

Cosine similarity model:
Open text responses from applicants were used to extract additional financial circumstances faced by each household. Natural language processing was used to extract these features and categorize them in one of six categories: education, housing,

divorce, medical, immigration, and employment. The number of circumstances faced by a household is multiplied by a set circumstance_cost value and added to the financial need sum for the household. A cosine similarity model was developed to determine if open text responses alone can predict funding eligibility. Text was normalized with lemmatization and vectorized with a tfdif vectorizer, calculating similarity between all text-sample pairs.

```
def stem_tokens(tokens):
    return [stemmer.stem(item) for item in tokens]

def normalize(text):
    return
stem_tokens(nltk.word_tokenize(text.lower().translate(remove_punctuation_map)))

vectorizer = TfidfVectorizer(tokenizer=normalize, stop_words='english')

def cosine_sim(text1, text2):
    tfidf = vectorizer.fit_transform([text1, text2])
    return ((tfidf * tfidf.T).A)[0,1]
```

Linear regression models:
Regression was used to understand how each feature, including parameter levels, influences funding allocated. Three regression models were tested for how well they explained feature importance: ordinary least squares (full and reduced model), poisson regression (reduced model), and ridge regression (reduced model).

Ordinary Least Squares

```
regr = linear_model.LinearRegression(random_state=42)
```

Poisson Linear Regression: For data with skewed distribution , such as our response.

```
poisson_regr = linear_model.PoissonRegressor(alpha=1.0, fit_intercept=True,
        max_iter=100, tol=0.0001, warm_start=False, verbose=0,  random_state=42)
```

Ridge Regression: Addresses models with multicollinearity. Regularization, alpha = 50.

```
ridge_regr = linear_model.Ridge(alpha=50.0, fit_intercept=True, copy_X=True,
        max_iter=None, tol=0.0001, solver='auto', positive=False,  random_state=42)
```

The analyst sets three parameter values: single_parent_cost (half or whole), circumstance_cost (low or mid), and disposable_unit (low, mid, or high) (Table 1). A fabricated dataset was created and all combinations of set parameters were applied to select the best combination of levels.

The response variable, the percent eligibility of each applicant, is binned into the five funding tiers. These are treated as continuous data in the model, as they occur along a range from 0 to 90.

Model feature definitions:
Response variable:
percent_eligible - the amount of scholarship award, by tier, that applicant is eligible for. Tiers = 90%, 80%, 70%, 50%, 0%

Applicant data:
household_income - Amount earned by adults in household.
number_adults - Number of additional adults in household is multiplied by estimated cost of an adult.
number_children - Number of children in household is multiplied by estimate cost of an additional child.
single - binary feature; 0 = not single parent, 1 = single parent
total_circumstances - Number of circumstances reported by family.

Features calculated using MIT values**:**
Base_income - Income required by a household of one adult and no children.
single parent value - Additional cost per child for single parent.
additional_adult - Cost of each additional child in household.
additional_child - Cost of each additional adult in household.

Parameters set by analyst:
single_parent_cost -Single parent additional need:
whole: full amount every child
*single_parent_cost = df_annual ['number_children'] * single_costdf_annual ['single']*
half: full amount first child and half for each additional.
*single_parent_cost = ((single_cost + (df_annual ['number_children'] - 1 (single_cost/2)) * df_annual ['single'])*
circumstance_cost - Number of circumstances multiplied by set parameter circumstance_cost level. As circumstance_cost increases, scholarship awards will increase.
low or mid
disposable_unit - Set parameter defining an amount of disposable income per child (full unit) and adult (half unit); family disposable income is divided by units to determine eligibility tier. As disposable_unit increases, scholarship awards will increase.
low, mid, or high

Engineered features**:**
extra_cost = *df_annual ['total_circumstances']* circumstance_cost + single_parent_cost*
family_base_income: income family needs to live up to standard
base + ((df_annual ['number_adults'] - 1) add_adult) + ((df_annual ['number_children']) add_child)
family_need_income = *family_base_income + extra_cost*
disposable_income = *household_income - family_need_income*
family_disposable_unit: a disposable unit of income as calculated based on family

structure and set disposable unit parameter.
(df_annual ['number_adults'] (disposable_unit/2)) + (df_annual['number_children'] disposable_unit)

Findings:
Exploratory data analysis:
Households of participants in BGS programming are primarily low-income, earning less than the baseline living wage estimated for Alameda County, California (Fig. 1). Almost 60 percent of housholds are headed by a single parent (Fig. 2). Education and housing are major additional costs for families in the BGS community Fig. 2). The manual approach to scholarship allocation in previous years resulted in highly uneven awards and some families receiving less funding than they would have been eligible for (Fig. 3).

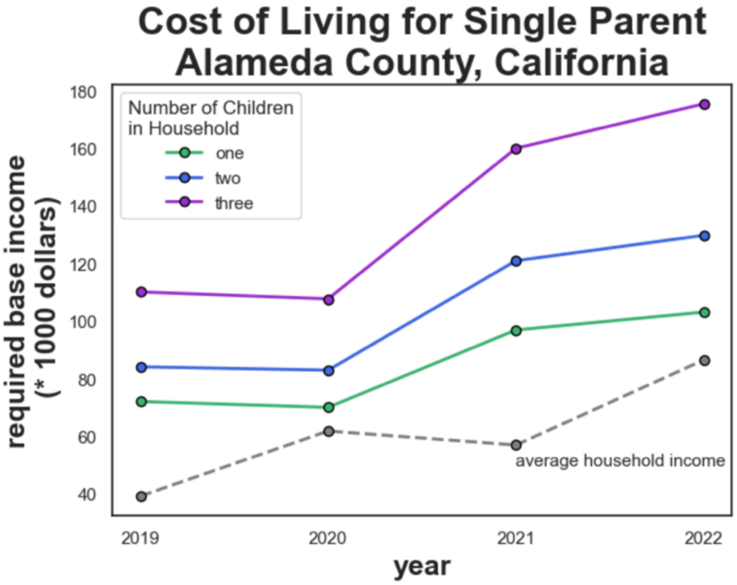


Fig. 1 Cost of living by year for single parent and one, two, or three children in Alameda County compared to average household income of participants (grey dashed line).

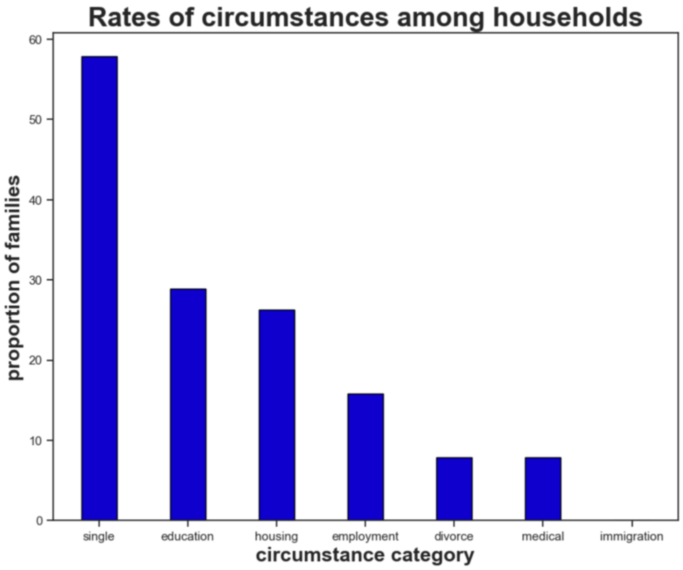


Fig. 2 Summary of additional circumstances faced by participant households.

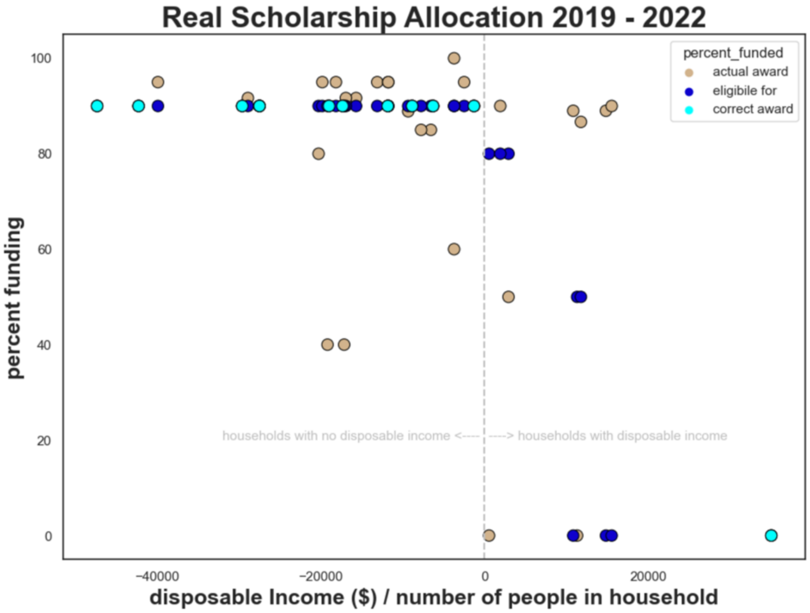


Fig. 3. Applicant funding (tan) is plotted with the award each participant was eligible for according to the tool (dark blue). Cases where the award amount matched the amount eligible for are marked as correct awards (aqua).

Cosine similarity model:
The text provided by applicants describing additional family circumstances was not useful for predicting scholarship award eligibility. Despite circumstances falling into six categories after natural language processing, text samples were not similar to one another, with over 83% of samples being less than one percent similar to another text sample. Samples with some similarity in text did not necessarily have similar award eligibility (Fig. 4).

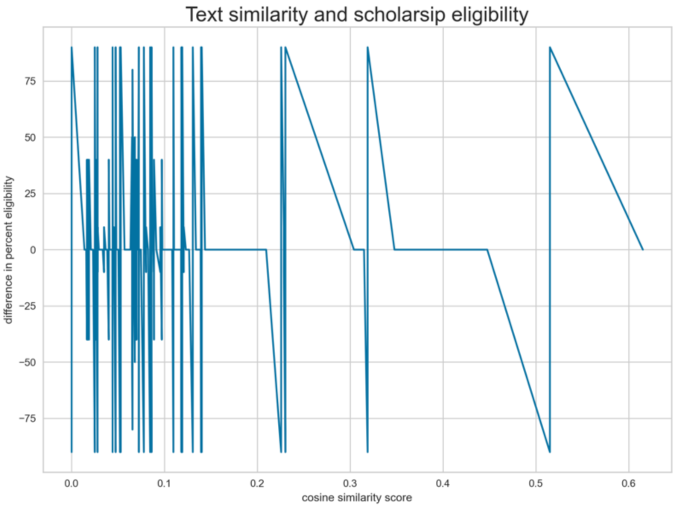


Fig. 4. Cosine similarity (x-axis) plotted with the difference in eligibility between paired observations (y-axis).

Linear regression models:

Linear regression method	MSE	r-squared training	r-squared testing
Ordinary least squares full	371.76	0.580	0.583
Ordinary least squares reduced	395.07	0.571.	0.557
Poisson	537.09	0.331	0.326
Ridge regression	394.96	0.571	0.557

Table 2. Model evaluation metrics comparison

Of the three regression models tested, the ridge regression model with regularization performed best. The ridge regression is a reduced model with twelve explanatory variables. The model was able to explain approximately 56% of the variation in the data (r-squared = 0.557) with a fairly high mean squared error (MSE = 394.96). This is only a slight improvement over the ordinary least squares reduced model. Plotted residuals are clumped into diagonals because of the binning effect of scholarship tiers, but are fairly evenly distributed around 0 (Fig. 5). The full model performed slightly better (r-squared = 0.58), but its interpretation is much more complicated with engineered values being related to other explanatory variables (Table 2).

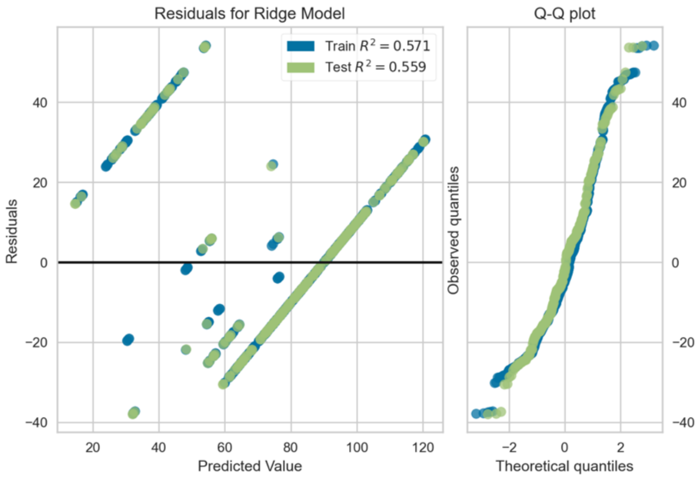


Fig. 5. Ridge regression residuals and Q-Q plots.

Household income has the largest effect on award; as income increases, award values decrease (Table 3). This is not surprising given that most applicants have incomes so low that they have no disposable income and get 90% funding. The number of children in the household and total_circumstances are positively related to the amount of award, and this is also intuitive. A household with more income and a greater number of additional financial circumstances will require more income and receive larger awards.

Other features had a fairly low influence on awards but we gather that, of the set parameters, the disposable_income_unit is most influential. Higher set values of disposable_income_unit leads to lower awards, which is intuitive (Table 3). Single_cost and circumstance_cost settings have a very small effect on awards in comparison to the major features, household_income, number_children, total_circumstances, and number_adults. Interestingly, being a single parent was associated with a lower award despite increasing the calculated financial need, but the effect was very small. The cost of having an additional adult may outweigh the additional need awarded to a single parent.

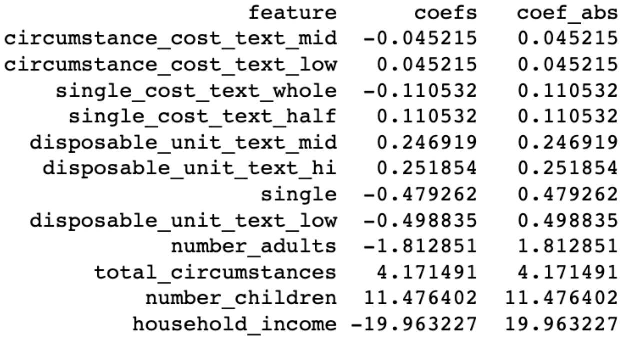

```
                   feature       coefs   coef_abs
circumstance_cost_text_mid   -0.045215   0.045215
circumstance_cost_text_low    0.045215   0.045215
    single_cost_text_whole   -0.110532   0.110532
     single_cost_text_half    0.110532   0.110532
  disposable_unit_text_mid    0.246919   0.246919
   disposable_unit_text_hi    0.251854   0.251854
                    single   -0.479262   0.479262
  disposable_unit_text_low   -0.498835   0.498835
             number_adults   -1.812851   1.812851
       total_circumstances    4.171491   4.171491
           number_children   11.476402  11.476402
          household_income  -19.963227  19.963227
```

Table 3. Feature importance in ridge regression (r-sqared = 55.7, MSE = 394.96. coef = coefficient value from model, coef_abs = absolute value of coefficient.

Next Steps:

The fabricated dataset will be expanded to, hopefully, improve regression model performance. The algorithm code will be updated each year to bring in the most recent MIT living wage data. In coming years, improvements will be made to the application format, eliminating the natural language processing part of the program.

Recommendations:
It is recommended to use the scholarship allocation tool to improve evenness and transparency in scholarship funding allocation. Use of the tool requires an annual update to the cost of living table but saves time in calculating award amounts as well as provides applicants with transparency around award criteria. An improvement to the application format was recommended. Applicants can simply check boxes for each of the six additional financial circumstances and eliminate the need for natural language processing of an open text entry. Adjusting the disposable_unit level will adjust scholarship awards up or down in an even approach across applicants. Adjusting circumstance_cost and single_parent_cost levels will impact only those applicants who have these conditions present; this should be done with caution. The allocation tool is a great asset for the organization and has improved the scholarship process.