ECON1000 - Using Big Data to Solve Economic and Social Problems
Economics and Watson
Brown University
Prof. John Friedman

BROWN

ECON 1000

Empirical Exercise #3 Submission Template

Due TUESDAY 10/10/2023, by 10am EDT on Gradescope

Gradescope Course Link: https://www.gradescope.com/courses/565275

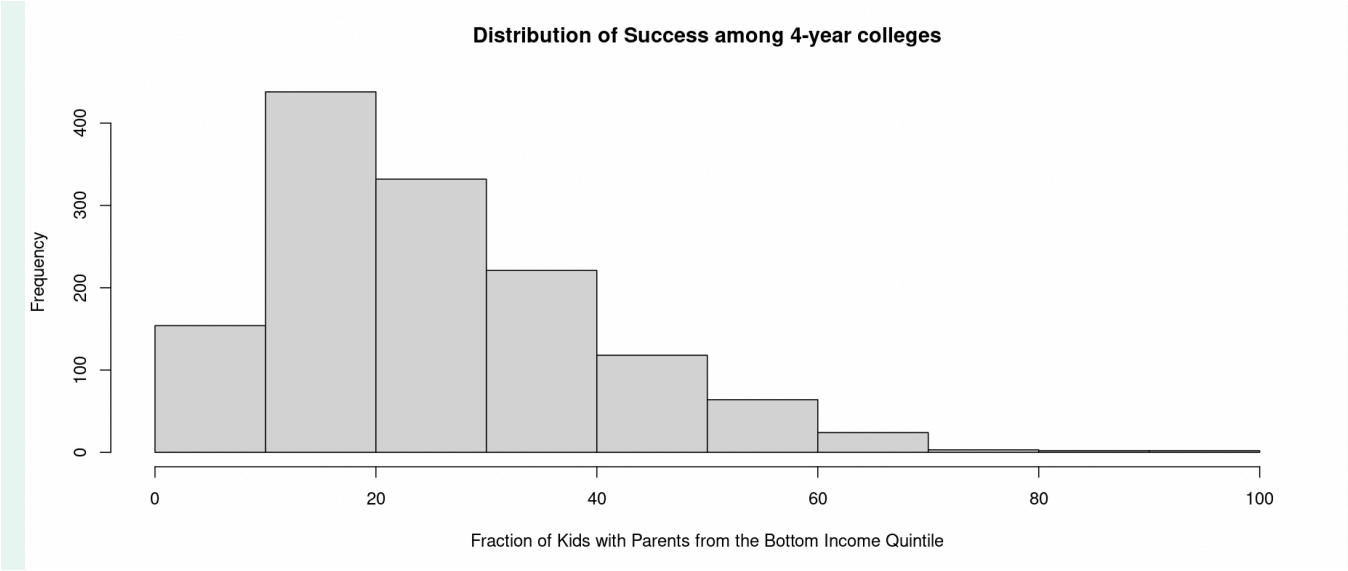
Name: Sonya Rashkoan

Group members with whom you worked[1]:

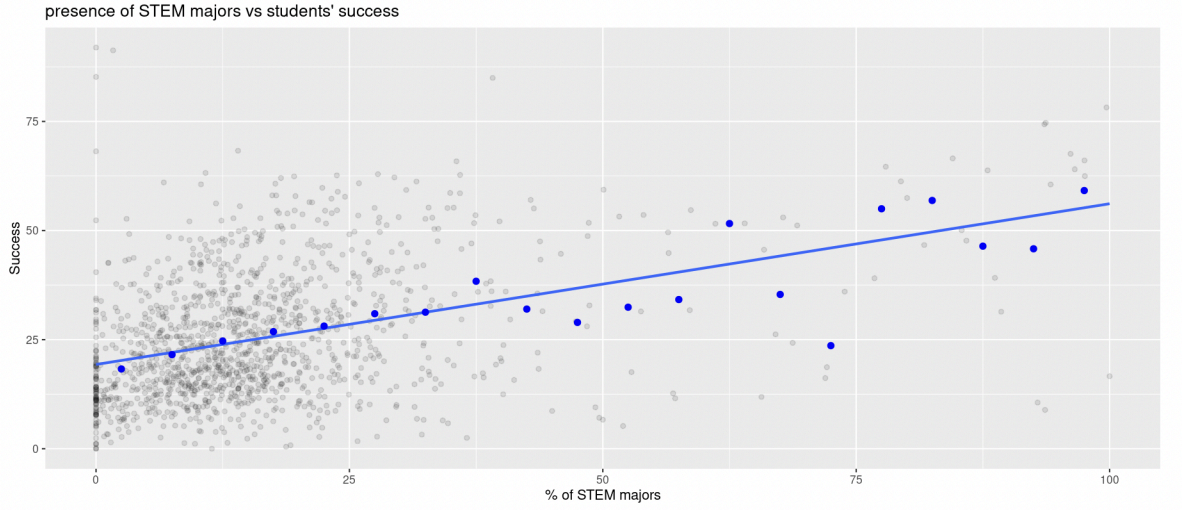
1. Seth Ables
2.
3.
4.

[1] In this class we encourage working in groups because you will learn a great deal from your peers. At the end of the day, however, it is important that you write up your own analysis. Concretely, this means you should choose your own city or county for the full exercise. Your narrative will be at the bottom of this document. Please then list all group members with whom you worked. If you have any questions, please ask.

1. Create a count-weighted histogram of success among 4 year schools.



2. Test whether colleges with a greater STEM focus have higher success rates among its students.



3. Estimate the best-fit line in your previous binscatter using a regression.

```
Call:
lm(formula = success ~ pct_stem, data = colleges_4yr)

Residuals:
    Min      1Q  Median      3Q     Max
-44.897  -9.137  -2.019   8.012  72.619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.31031    0.53246   36.27   <2e-16 ***
pct_stem     0.36830    0.02436   15.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.4 on 1353 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.1446,    Adjusted R-squared:  0.1439
F-statistic: 228.6 on 1 and 1353 DF,  p-value: < 2.2e-16
```

Coefficient: positive thus it indicates that as the share of STEM students increases, the success rate of students also increases

Standard error: the standard error is very low which indicates that the calculated mean is an accurate reflection of the actual population and can be taken for its truth value

P-value: is much lower than 0.05, we can conclude that the variable affects the dependent variable aka presence of STEM majors at a college leads to that college’s student success

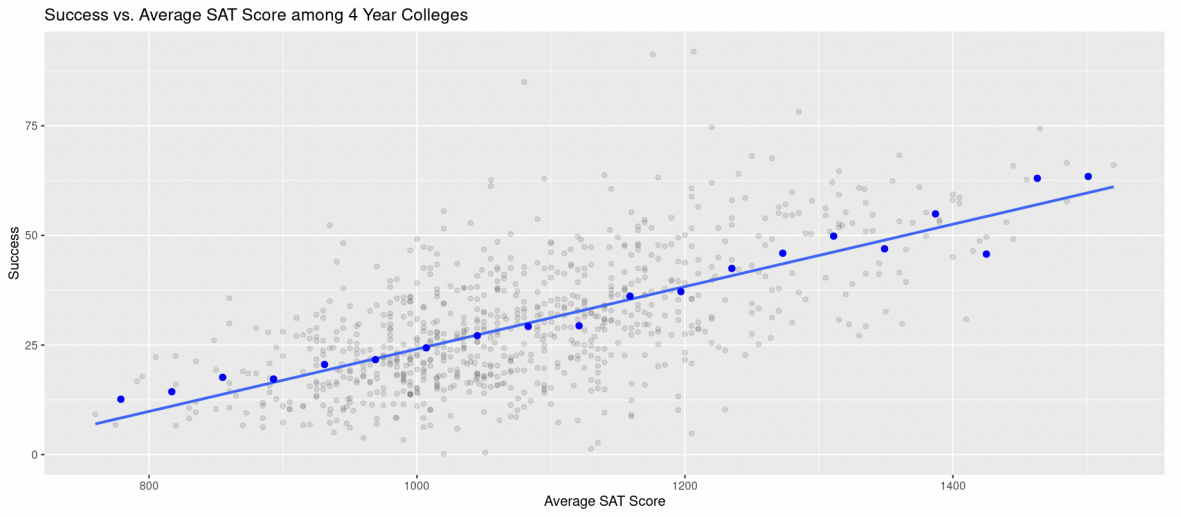
4. Test whether more selective colleges have higher success rates among its students.

```
Call:
lm(formula = success ~ sat_avg, data = colleges_4yr)

Residuals:
    Min      1Q  Median      3Q     Max
-33.874  -7.691  -0.835   7.158  55.186

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -47.057611   3.396883  -13.85   <2e-16 ***
sat_avg       0.071154   0.003117   22.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.42 on 810 degrees of freedom
  (546 observations deleted due to missingness)
Multiple R-squared:  0.3914,    Adjusted R-squared:  0.3907
F-statistic:   521 on 1 and 810 DF,  p-value: < 2.2e-16
```



Coefficient: is positive thus it indicates that there is a positive relationship between students' SAT scores and their success

Standard error: is low which indicates that the calculated mean is an accurate reflection of the actual population and can be taken for its truth value

p-value: is much lower than 0.05, so we can conclude that the variable affects the dependent variable aka average SAT score which determines the college's selectivity during admissions leads to that college's students' higher success – doesn't mean that the correlation is necessarily true but affirms the coefficient for this particular case

5. Run a multivariate regression to test between STEM focus and selectively in predicting student success.

```
Call:
lm(formula = success ~ pct_stem + sat_avg, data = colleges_4yr)

Residuals:
    Min      1Q  Median      3Q     Max
-33.780  -7.398  -0.928   6.644  59.137

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.837684   3.401667 -11.711  < 2e-16 ***
pct_stem      0.226195   0.028726   7.874  1.1e-14 ***
sat_avg       0.060896   0.003276  18.587  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.01 on 809 degrees of freedom
  (546 observations deleted due to missingness)
Multiple R-squared:  0.4348,    Adjusted R-squared:  0.4334
F-statistic: 311.1 on 2 and 809 DF,  p-value: < 2.2e-16
```

Coefficient: is positive and thus indicates that % STEM majors and average SAT score at a college lead to students' success

Standard error: is very low for both which indicates the calculated mean is an accurate reflection of the actual population and can be taken for its truth value

P-value: is lower than 0.05 thus showing that both the % STEM majors and average SAT score affect the students' success rate – doesn't mean that the correlation is necessarily true but affirms the coefficient for this particular case

6. Choose a different measure of success and a coherent subset of colleges. Explore the predictors of your new success in these schools using multivariate regression.

Success of students based on college sticker price and female student population in 4-year colleges in the state of NY

```
Call:
lm(formula = success ~ sticker_price + female, data = colleges_4yr)

Residuals:
    Min      1Q  Median      3Q     Max
-47.046  -8.625  -1.709   8.431  65.005

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.16414    1.63770  16.587  < 2e-16 ***
sticker_price  0.86129    0.05166  16.672  < 2e-16 ***
female        -0.19163    0.02739  -6.997 4.21e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.91 on 1269 degrees of freedom
  (86 observations deleted due to missingness)
Multiple R-squared:  0.1954,    Adjusted R-squared:  0.1941
F-statistic: 154.1 on 2 and 1269 DF,  p-value: < 2.2e-16
```

Coefficient: is positive for the relationship between sticker price and success, meaning that the more expensive the college gets, the more successful the students are after college. The coefficient is negative for the female student population indicating that female student population and students' success rate move in different directions

Standard error: is low for both variables and thus the calculated mean is an accurate reflection of the actual population and can be taken for its truth value

P-value: is lower than 0.05 thus showing that the sticker price of the college and female student population affect the college's access opportunities for low-income students – doesn't mean that the correlation is necessarily true but affirms the coefficient for this particular case

Please submit this entire document to Gradescope. **You must select the corresponding pages for each answer when uploading this document to Gradescope.** Failure to do so makes it difficult to locate your answers.