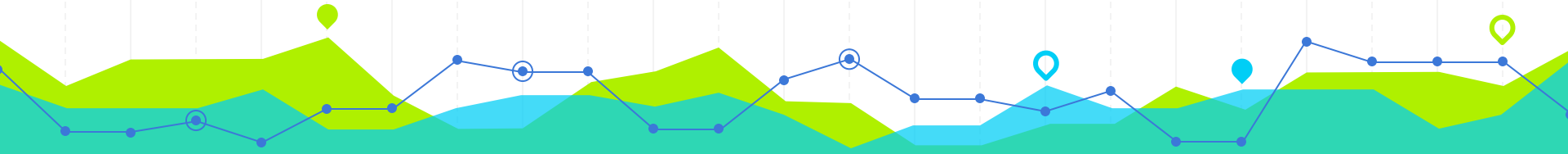
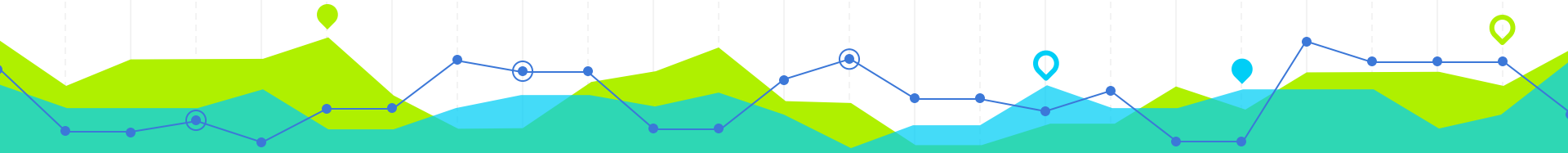


Analyzing SOCT and Grade Data for MSU Courses

BY WILLIAM CHETTLEBURGH

CMSE 202 - 002



Section 1: Introduction

A brief background and overview of techniques

Background on Datasets

- After each course, MSU students complete an SOCT (Student Opinion of Courses and Teaching) survey, which includes the questions:
  1. Overall was the instructor effective?
  2. Overall was the course worthwhile?
  3. Was the instructor available to help students outside of class?
  4. Was the workload reasonable for the mastery of the course material?
  5. Was the course well organized?
  6. Was I interested in taking this course?
- Reports of the grades received by course are available at msugrades.com:
  - Obtained through the Freedom of Information Act
  - Includes counts of each grade received on the 4.0 scale

Project Questions

Using these datasets for the Spring 2020 semester, this project seeks to answer:

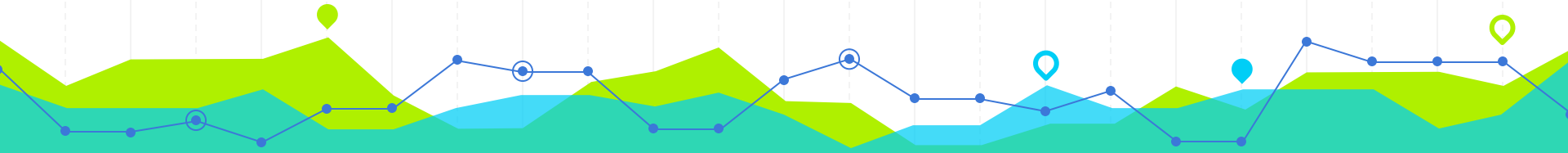
1. Is there a correlation between the responses students provide on SOCT survey questions and the grades they receive?
2. How do the SOCT survey responses vary between subjects? Do certain subjects receive more favorable responses, on average?
3. How do the SOCT survey responses and average grades vary between course levels?

Models Used

- The first question can be answered through regression:
  - Linear regression is used to determine statistical significance (p-values) and practical significance (magnitude of coefficients)
  - Multivariable regression can be used to understand the effect of a single variable *while keeping the others constant*
  - Mathematically, this model finds a line that minimizes the sum of the squared residuals (or errors). The p-value is calculated through statistical inference with the t-distribution (a p-value below 0.05 is considered significant)
- The last two questions are answered through visual models:
  - Boxplots are used for comparisons of center (median) and spread (IQR)
  - Ridgeplots are used for more careful analysis of the distributions

Overview of Computational Techniques

- Selenium is used for performing the webscraping of the SOCT website
- Pandas is used to transform and merge the two datasets
- Matplotlib, Seaborn, and JoyPy are used to visualize the results (with scatter/density plots, boxplots, and ridgeplots respectively)
- Statsmodels is used to perform regression and analyze the coefficients/p-values
- OLSplots is used to analyze whether the conditions for linear regression are met
- Boxcox from scipy is used to normalize the skewed distributions (this is briefly covered in the conclusion)



Section 2: Data Collection

A demonstration of webscraping and merging

SOCT Website Interface

Select subject

(Subject codes for S.O.C.T. forms received since Spring 2001 semester.)

CMSE - Comp Math,Science,Engineering Submit

Courses

Listed below are all course numbers within this subject for SOCT forms received since Spring 2001 semester

Select Course

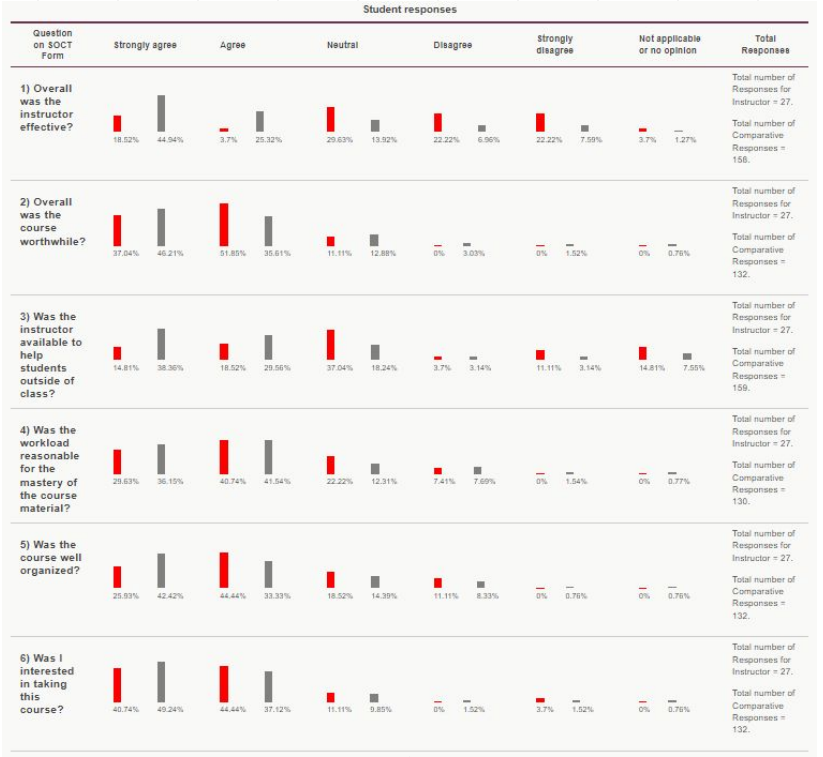
202 - Comp Model & Data Anyl II Submit

Instructor

The dropdown box below lists all instructors who have taught this course within the last two years and for whom SOCT forms have been received.

Select Instructor

Min Chen Submit



Under the Hood: HTML

```
▼<div class="yui-u">
  <h1>Select a rating to view</h1>
 ▼<form name="form_select_subject" action="Main_Page.asp?Page=Select_Course" method="POST">
  ▶<label for="Select_SubjectCode">…</label>
  ▼<select name="Select_SubjectCode" id="Select_SubjectCode">
     <option value="AAAS">AAAS  -  African American & Afr Studies</option>
     <option value="AAE">AAE  -  Advanced Academic English</option>
     <option value="ABM">ABM  -  Agribusiness Management</option>
     <option value="ACC">ACC  -  Accounting</option>
     <option value="ACM">ACM  -  Arts and Cultural Management</option>
     <option value="ACR">ACR  -  Comm Ag Rec & Res Studies</option>
     <option value="ADV">ADV  -  Advertising</option>
     <option value="AE">AE  -  Agricultural Engineering</option>
     <option value="AEE">AEE  -  ANR Edu & Com Sys</option>
     <option value="AESC">AESC  -  Applied Engineering Sciences</option>
     <option value="AFR">AFR  -  African Languages</option>
     <option value="AIIS">AIIS  -  Amer Indian & Indigenous Stdy</option>
     <option value="AL">AL  -  Arts & Letters</option>
     <option value="AMS">AMS  -  American Studies</option>
     <option value="ANP">ANP  -  Anthropology</option>
     <option value="ANR">ANR  -  Agriculture & Nat Resources</option>
     <option value="ANS">ANS  -  Animal Science</option>
     <option value="ANTR">ANTR  -  Human Anatomy</option>
     <option value="ARB">ARB  -  Arabic</option>
     <option value="AS">AS  -  Aerospace Studies</option>
     <option value="ASC">ASC  -  Audiology and Speech Sciences</option>
     <option value="ASN">ASN  -  Asian Languages</option>
```

The menu options are given within <select> and <option>. In a loop, they can be selected and submitted with Selenium.

```
▼<table id="student-responses">
   <caption>Student responses</caption>
  ▶<thead>…</thead>
  ▼<tbody>
   ▼<tr>
    ▶<td class="response-question">…</td>
    ▼<td class="bar-cell text-center">
     ▼<table class="table-graph">
      ▼<tbody>
       ▼<tr>
        ▼<td>
           <img src="images/bar1.jpg" alt width="10" height="20">
          ▼<p class="text-85">
             "18.52% "
             <span class="hide">of total student responses for this instructor</span>
            </p>
           </td>
          ▶<td>…</td>
          </tr>
         </tbody>
        </table>
       </td>
      ▶<td class="bar-cell text-center">…</td>
      ▶<td class="bar-cell text-center">…</td>
      ▶<td class="bar-cell text-center">…</td>
      ▶<td class="bar-cell text-center">…</td>
      ▶<td class="bar-cell text-center">…</td>
      ▶<td>…</td>
```

The SOCT data is given in a table. The data can be accessed with the find_elements function (and parsed using string functions such as split).

Comparison of SOCT Data and MSU Grades Data

	subject	subject_desc	course	prof	effective_prof	worthwhile_course	help_available	workload	course_organization	course_interest
0	AAE	Advanced Academic English	220	Andrew S McCullough	0.875013	0.854175	0.833325	0.791675	0.750000	0.791675
1	AAE	Advanced Academic English	220	David Krise	0.937500	0.906250	0.906250	0.906250	0.906250	0.937500
2	AAE	Advanced Academic English	220	Laura Marian Ramm-Christensen	0.931825	0.931825	0.931825	0.909100	0.925000	0.909100
3	AAE	Advanced Academic English	221	Carol Elaine Arnold	0.916675	0.888900	0.972225	0.944450	0.972225	0.944450
4	AAE	Advanced Academic English	222	Carol Wilson-Duffy	0.937500	0.906250	0.937500	0.937500	0.937500	0.890625
...	...	...	...	...	...	...	...	...	...	...
2515	WS	Women's Studies	301	Laura Jean Apol	0.821457	0.839325	0.857175	0.803625	0.821450	0.839325
2516	WS	Women's Studies	304	Hillery Glasby	0.980775	0.942300	0.979173	0.884625	0.942300	0.903850
2517	WS	Women's Studies	403	Yuanfang Dai	0.825000	0.837500	0.881579	0.812500	0.787500	0.812500
2518	WS	Women's Studies	424	Kristin Mahoney	0.950000	0.950000	1.000000	0.900000	0.950000	0.950000
2519	WS	Women's Studies	492	Aminda Moine Smith	0.800000	0.850000	0.950000	0.950000	0.750000	0.900000

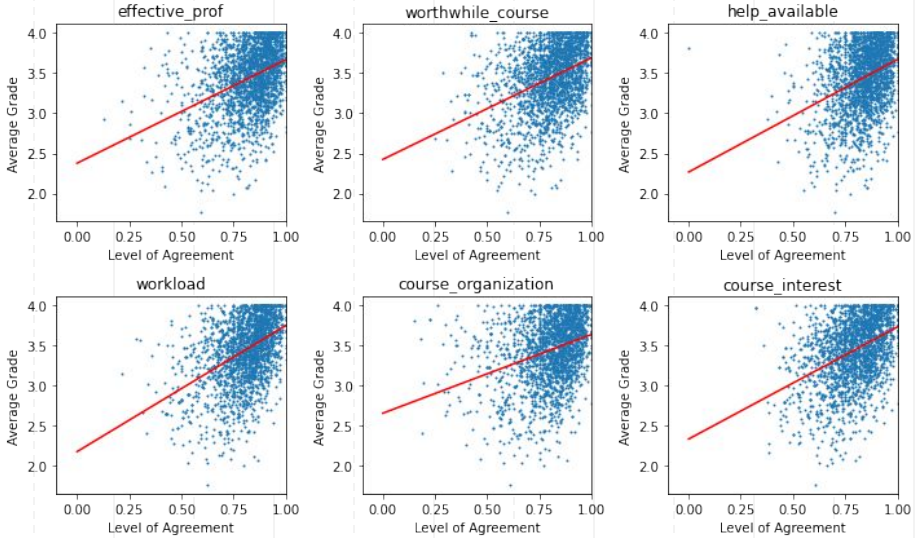
	subj_code	crse_code	Instructor	avg_grade
0	AAAS	100	CHAMBERS JR,GLENN A	3.647059
1	AAE	220	KRISE,DAVID	2.900000
2	AAE	220	MCCULLOUGH,ANDREW S	3.156250
3	AAE	220	RAMM,LAURA	2.937500
4	AAE	220	WALTERS,PATRICIA G	3.000000
...	...	...	...	...
3538	WS	301	APOL,LAURA J	3.976190
3539	WS	304	GLASBY,HILLERY	3.826087
3540	WS	403	DAI,YUANFANG	3.769231
3541	WS	424	MAHONEY,KRISTIN	3.538462
3542	WS	492	SMITH,AMINDA M	3.750000

- Above are the two datasets after performing weighted averages on the data
- The two instructor columns are very different, preventing a simple merge from working
- Instead, the datasets are merged/grouped on subject and course, and then a “similarity score” counting the number of matching components in the names are calculated within each group. The row with the highest (and nonzero) similarity score is kept.

Section 3: Correlations

Finding the relationships between SOCT responses and average grades

Analysis of Single-Variable Regression Models



	rsquared	intercept	coeff	coeff_pvalue
effective_prof	0.153260	2.378232	1.287781	9.924464e-86
worthwhile_course	0.152282	2.425553	1.264573	3.787748e-85
help_available	0.119121	2.267982	1.401410	8.329608e-66
workload	0.200749	2.175761	1.570262	8.384617e-115
course_organization	0.097976	2.654085	0.981408	7.717241e-54
course_interest	0.169900	2.332620	1.394963	9.938801e-96

All the predictors are statistically significant (low p-value), but some are more practically significant (high magnitude of coefficient)

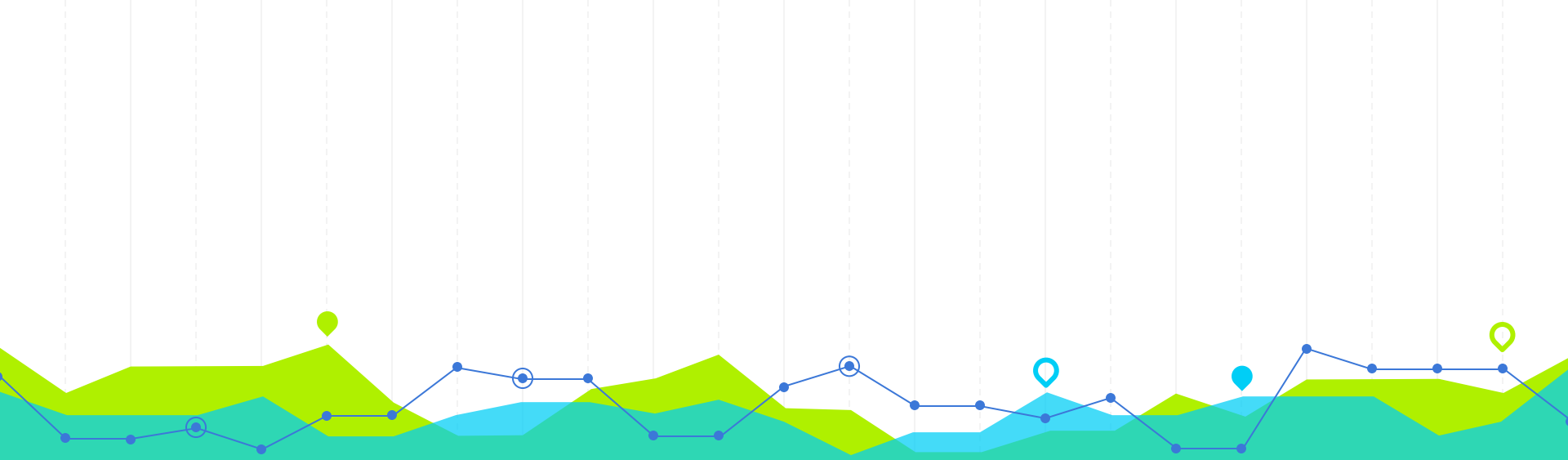
Analysis with Multivariable Regression

Dep. Variable:	avg_grade	R-squared:	0.248
Model:	OLS	Adj. R-squared:	0.247
Method:	Least Squares	F-statistic:	127.3
Date:	Sun, 05 Dec 2021	Prob (F-statistic):	2.19e-139
Time:	21:59:55	Log-Likelihood:	-837.28
No. Observations:	2317	AIC:	1689.
Df Residuals:	2310	BIC:	1729.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0139	0.066	30.506	0.000	1.884	2.143
effective_prof	0.7405	0.161	4.586	0.000	0.424	1.057
worthwhile_course	-0.6073	0.180	-3.377	0.001	-0.960	-0.255
help_available	-0.0429	0.143	-0.301	0.764	-0.323	0.237
workload	1.8101	0.142	12.784	0.000	1.532	2.088
course_organization	-1.0325	0.127	-8.116	0.000	-1.282	-0.783
course_interest	0.8931	0.128	6.991	0.000	0.643	1.144

	Variance Inflation Factors
effective_prof	342.384590
worthwhile_course	402.205559
help_available	200.062404
workload	248.708956
course_organization	206.087782
course_interest	191.432018

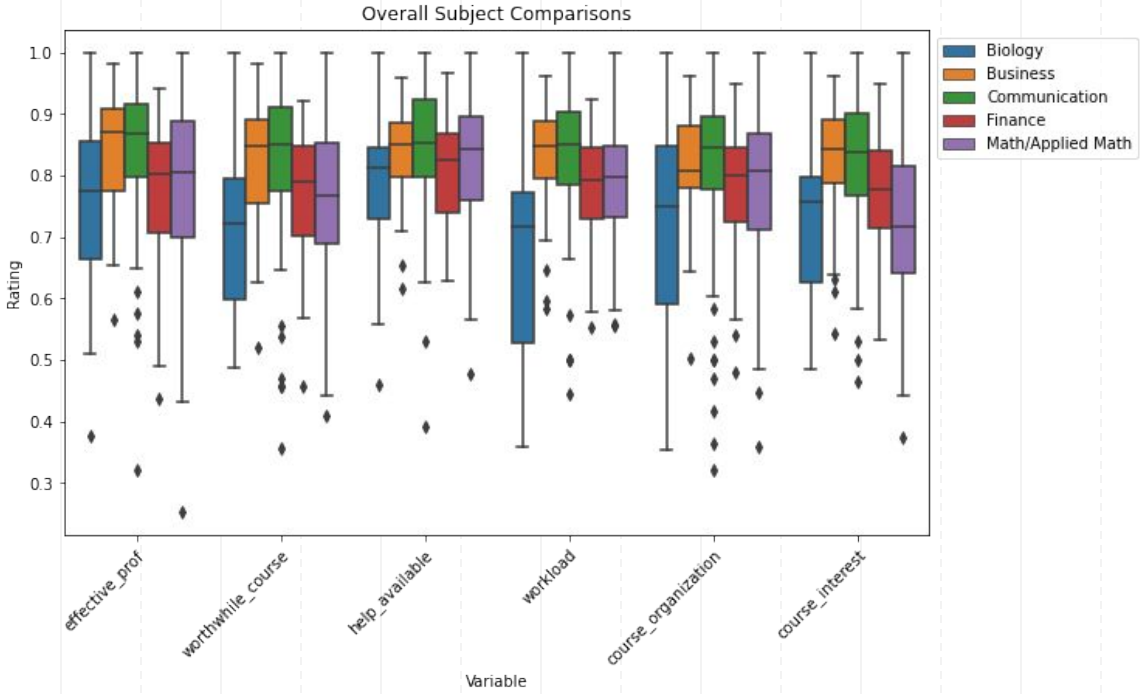
The results for the coefficients are quite unexpected. To explain this, we note that the variance inflation factors are large (above 10), indicating that the predictors are strongly interrelated. Thus, collinearity is causing the model to be inaccurate.



Section 4: Comparisons

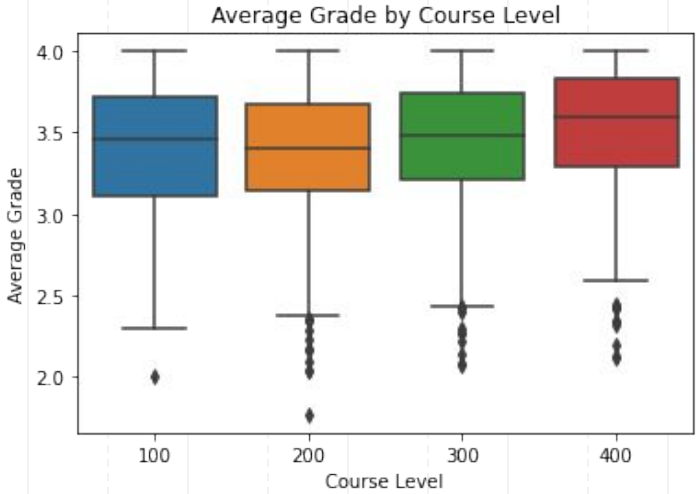
Comparing SOCT responses and average grades for subjects and course levels

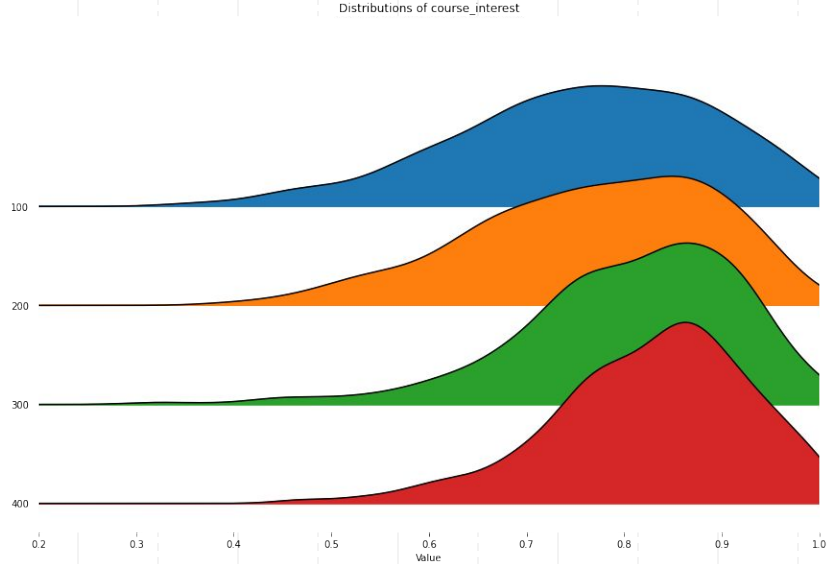
SOCT Responses by Subject Category



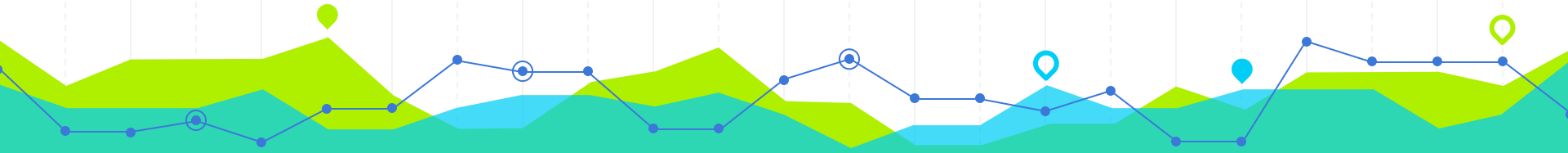
- Business and Communication classes tend to be ranked well
- Finance and Mathematics are ranked moderately
- Biology is ranked the worst (except in course interest, where mathematics falls lower)

Comparisons by Course Level





- There is possibly a small increase in average grade as course level increases
- Course interest increases as course level increases (center increases and spread decreases)



Section 5: Conclusions

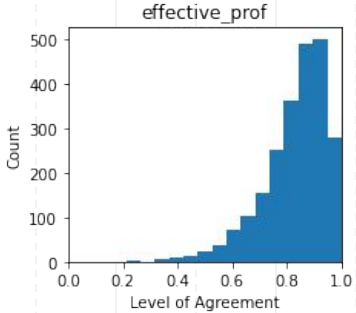
Summary of findings and difficulties encountered during the project

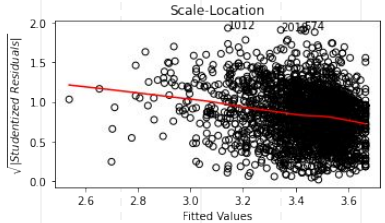
Recap of Findings

- All of the SOCT responses have a significant positive correlation with average grade
  - Reasonability of workload and help availability have the largest impacts, given by their coefficients
  - All the responses are strongly correlated with each other
- Business and Communication classes tend to be ranked better than Finance and Mathematics, which is in turn better than Biology
- There is a possible increase in average grade for higher course levels
- The course interest tends to increase as course level increases

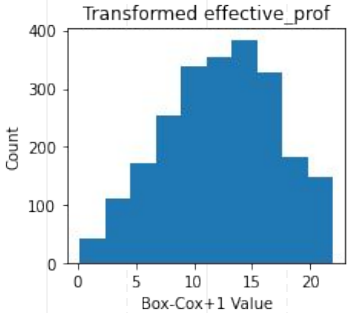
Difficulties Encountered During the Project

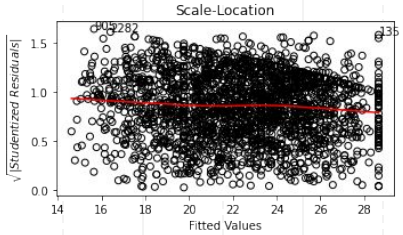
- Merging the data was very difficult. Several cases had to be checked by hand due to different spellings, maiden/married names, etc.
- The conditions for linear regression were not always strongly met. Through Box-Cox transformations, several of these were fixed, although this resulted in models that could not be directly compared.

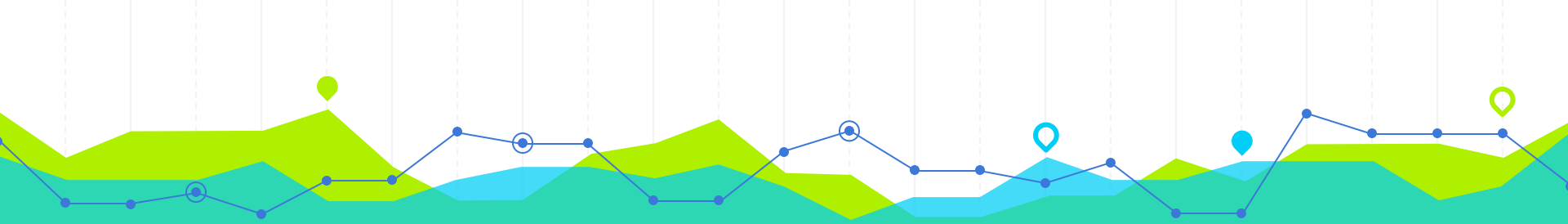




$$y = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln x & \text{if } \lambda = 0 \end{cases}$$







Thanks for reading! If you have any questions, or would like access to the code, contact chettleburghw@gmail.com.

Data From:

- https://soct.msu.edu/
- https://msugrades.com/

External Files and Code From:

- https://github.com/j-sadowski/FromRtoPython/blob/master/OLSplots.py
- https://chromedriver.chromium.org/

Other Sources:

- https://towardsdatascience.com/everything-you-need-to-know-about-multicollinearity-2f21f082d6dc
- https://www.usnews.com/best-colleges/michigan-state-2290/academics
- https://stackoverflow.com/a/44713292
- https://stackoverflow.com/a/60215826
- https://towardsdatascience.com/ridgeline-plots-the-perfect-way-to-visualize-data-distributions-with-python-de99a5493052