Mental Health Issues in STEM Fields

June 12, 2019

1 COGS108 - Final Project

2 Overview

In this project, we first came up with our research question that if the academic and social factors contribute to STEM students' mental illness. Many STEM students who go out into the workplace struggle with mental health conditions that can begin to develop in college. We found data sets online, cleaned them and did several data visualizations to analyze and made our conclusions. We have data that shows students' grades, the number of hours studied per week, employment rates after after college, and answers to questions on mental health in the workplace.

3 Names

- Alessandra Landingin (A11936362)
- Bettina Gerez (A13718015)
- Chau Dang (A13808198)
- Duo(Cassie) Yu (A13459470)
- Jiayu Fu (A15697380)
- Rachel Lim (A14514205)

4 Research Question

What academic and social factors do STEM (Specifically Technology, Engineering, and Mathematics) students experience during undergrad that contribute to mental-health illnesses in their professional careers?

5 Background and Prior Work

In the world we live in today, with everyone striving to work at top-tier companies and earn as much money as possible to support themselves, it is not surprising that the world we live in today is one filled with competition. One of the more competitive fields in the STEM field, and a position in this field can be very prestigious and well-paying if you do it right. However, with so many people and recent graduates trying to establish themselves in the industry, it can be very difficult to get into those positions. As a result, people push themselves really hard, as

early as college or high school, so that they can excel in their field of study. This can cause people to become overwhelmingly stressed, and it can be very unhealthy both physically and mentally. This often causes students to struggle with mental health issues. According to this article http://bit.ly/2veT8Vv, not much is known about the academic factors that contribute to the mental health of students since there are many other factors that should be taken into account, such as demographic and social factors. The article states that male students are more likely to be at risk for suicide, while female students are more likely to have major depression and anxiety disorders, and students from lower socioeconomic backgrounds tend to have more anxiety as well. Therefore, if we want to get accurate data representations, we should make sure to include students of all different types in our sample, or we should analyze students with different traits separately (ex: analyze female students of similar racial background together, and analyze male students of similar socioeconomic background together, etc.) and then compare them. Things such as striving for perfectionism definitely play a part in the amount of stress that students report. This article, https://journals.sagepub.com/doi/10.1177/0146167204272298, states in its abstract that perfectionism can be self-imposed or it can be dictated by peer pressure. Those with non-self-determined academic motivation experience higher levels of psychological adjustment difficulties than those with self-determined academic motivation do. Aside from an individual's goal of perfectionism, there are barriers that are cultivated within the STEM community. In "Barriers and Opportunities for 2-Year and 4-Year STEM Degrees: Systemic Change to Support Students' Diverse Pathways", it discusses the notion that educators tend to be particular about what counts as "scientific reasoning and sense-making." This, in turn, leads to the barrier that students may feel between themselves and their professors. It becomes a problem when they become too afraid to ask questions about topics or subjects that they might need help on. Then, the student becomes overwhelmed with the rigorous subjects and ultimately leads to the deterioration of their mental health condition. That is why statistics have shown that about 15% of college students suffer anxiety, and this causes them to have difficulty functioning in academic settings. (Reinberg 2018) This deteriorated mental health condition, if left untreated, will eventually bleed into their professional life. It is imperative that research is done to determine the factors that contribute to mental health, such that attempts are implemented to mitigate those factors as soon as they arise. Just as other health issues, like high blood pressure or diabetes, there are preventative measures that can be put in place to reduce or even eradicate any possible long-term effects of what is experienced during undergrad on the mental health of an individual in their professional career.

6 Hypothesis

Due to the rigorous and competitive nature of the undergraduate STEM curriculum, students in these majors typically experience a higher number of study hours per week, are more prone to backlogging, and show extreme distress about how their grades are in comparison to their peers. Eventually, these factors take a major toll on a student's confidence and self-worth in their foreseeable professional life, which ultimately leads to a range of mental health issues in STEM-related industries. This is because the need to do a lot of studying and be competitive places a lot of stress on people.

7 Dataset(s)

To relate mental health issues in STEM-related industries to how students performe at school, we are using these datasets

Exporatory Analysis on Worst Grades: This dataset is from the University of Washington, An Exploratory Analysis on Worst Grades from 2006 to 2017 found at The data has totals of 9,000 courses, almost 200,000 course sections, 3 million grades reported and 18,000 instructors. The dataset analyzed on which courses have the most students with failing grades and which professors are tough in these particular courses. With over 1,000 samples in each observation, the data narrows down the results of topmost 20 courses in the university with 14 courses being Mathematics. In this case, we can make the connection how often Math courses have brought students' grades down from failing the class – in which, it could lead them to doubt themselves pursuing a career in the STEM fields. Additionally, the data also includes which day of the week had the most passing and failing grades. We can factor out how it negatively affects the student's performance on certain courses. https://www.kaggle.com/mohitjoshi29/an-exploratory-analysis-on-worst-grades.

Mental Health in Tech Survy: This dataset is from a Mental Health in Tech survey from Kaggle.This is data was collected in 2014 from the OSMI Mental Health Survey. The survey collected about 1257 people's responses from all around the world. The survey is still ongoing and is still collecting more responses from employees in the tech industry. Each survey contains questions about how their mental well-being is being treated similarly to their physical well-being. https://www.kaggle.com/ashwinireddy/mental-health-in-tech-survey-rpart

CAPE Data: We obtained the CAPE data from cape.ucsd.edu. We downloaded the html from the websites and wrote a Java program to parse and retrieve the data from the html source code. We got the average number of hours studied and average grade received for each offering of the classes we selected, which were core classes in the different majors.

FiveThirtyEight College Majors: The third dataset we are using is found on Kaggle at https://www.kaggle.com/fivethirtyeight/fivethirtyeight-college-majors-dataset#grad-students.csv This is a dataset from FiveThirtyEight hosted on their Github. The data includes 173 unique majors with the specific graduate sample size for each of the major. This data set collects a lot of information related to the employment and unemployment of graduate and non-graduate students. Analyzing this data set, we can thereby conclude that the low employment rate might have some influence on students grades and can also put more pressure on the students in STEM majors.

These are the libraries we will be using to help us analyze the dataset

In [183]:
```
#Import

        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
```

8 Setup

In [184]:
```
#Load files into PandaFrame Exporatory_Analysis_on_Worst_Grades
course_offerings = pd.read_csv('Exporatory_Analysis_on_Worst_Grades/course_offerings
grade_dis = pd.read_csv('Exporatory_Analysis_on_Worst_Grades/grade_distributions.csv
```

In [185]:
```
#uploading csv file for the mental health in tech kaggle page
df_survey=pd.read_csv(open('Mental_Health_in_tech/survey.csv'))
```

```
In [186]: df_grad=pd.read_csv(open('FiveThirtyEight_Cassie/grad-students.csv'))

In [187]: #create dataframe from CAPE data
          capedf = pd.read_csv('capeData.csv')

          #create dataframe from non-STEM CAPE data
          capedf_nonstem = pd.read_csv('nonstem_capeData.csv')
```

9 Data Cleaning

We first clean up the dataset we are using the University of Washington, An Exploratory Analysis on Worst Grades from 2006 to 2017. We want to find out how do students performe in each course so we only keep courses offering, number of letter grade students recieve, total number of students and the rate of A and F by dropping uneccessary columns. To make our data more valid, we also dropped rows in which student numbers are less than 100 which is relative small. The second dataset we cleaned is the mental health in tech survy. Considering that we are only going to focus on mental health issues, so we dropped columns that is not useful to show menatl health issues such as physical health consequences and interviews.

```
In [188]: #Drop uneccessary columns
          course_offerings = course_offerings.drop(['term_code', 'course_uuid'],axis=1)
          grade_dis["students count"] = grade_dis.iloc[:,2:18].sum(axis=1)
          grade_dis = grade_dis.drop(['section_number','ab_count', 'bc_count', 's_count', 'u_c
                                      'n_count', 'p_count', 'i_count','nw_count', 'nr_count',
```

We calculate the A rate and D/F rate for each class and drop the class with total sudents less than 100.

```
In [189]: #Merging course_offerings and grade_dis on course_offerings_uuid and uuid
          courses_combined = pd.merge(course_offerings, grade_dis.set_index('course_offering_u
                                      left_on='uuid', right_index=True)

          #Drop any missing value in the 'name' row
          courses_combined = courses_combined.dropna(how='any')

          #Drop the ID column
          courses_combined = courses_combined.iloc[:,1:8]

          #Rename the columns
          courses_combined.columns = ["course", "A", "B", "C", "D", "F", "Total students"]

          # merge courses with the same name
          course_unique = courses_combined.groupby('course', as_index=False).sum()

          # drop the class with total students < 100 since thats relative small data
          course_unique = course_unique[course_unique['Total students'] > 100]

          #calculate the a rate for each course
```

```
course_unique["A rate"] = course_unique["A"] / course_unique["Total students"]

#calculate the DF rate for each course
course_unique["DF rate"] = ( course_unique["D"] + course_unique["F"])/ course_unique
course_unique
```

Out[189]:

	course	A	B	C	D	F	Total students	\
7	19&20 C Russian Lit Tran I	440	74	17	8	10	755	
8	19&20 C Russian Lit Tran II	305	45	12	3	8	511	
9	19&20th C Russn Lit Tran I	219	87	21	11	11	546	
10	19&20th C Russn Lit Tran II	157	62	10	0	5	364	
11	19th C Painting in Europe	46	17	3	5	1	175	
23	1st Semester Polish	71	7	5	1	0	111	
24	1st Semester Portuguese	293	160	113	40	17	818	
25	1st Semester Russian	415	138	33	18	17	923	
26	1st Semester Serbo-Croatian	75	10	0	0	2	107	
27	1st Yr Classical Chinese	226	90	26	4	2	567	
28	1st-Yr Seminar: Biological Sci	257	11	2	2	0	291	
29	1st-Yr Seminar: Humanities	191	82	9	4	5	447	
32	1st-Yr Seminar: Social Sci	309	49	6	0	1	460	
34	1st-Yr Smr: Soc Sci, Ethnic St	112	3	0	0	0	120	
39	20th Century Art in Europe	55	61	12	4	2	250	
41	20th Century Literature	111	5	0	0	0	147	
48	2nd Semester Portuguese	160	106	69	26	3	483	
49	2nd Semester Russian	258	86	16	4	7	508	
52	3D Digital Studio I	83	18	3	3	1	144	
60	3rd Semester Portuguese	147	78	42	8	5	346	
62	3rd Year Obs & Gynecology	391	538	3	0	1	1541	
63	3rd Year Pediatrics	359	343	0	0	0	1318	
64	3rd Yr Conversatn & Compositn	99	25	7	2	0	202	
65	3rd Yr Primary Care Clrkshp	335	304	1	0	0	1075	
72	4th Semester Portuguese	48	19	15	2	0	125	
74	4th Yr Compositn & Conversatn	154	32	3	1	0	256	
79	5th Semester Japanese	152	72	15	3	1	395	
86	A History of Greek Civ	166	72	10	5	5	425	
87	A History of Rome	392	281	40	9	10	1403	
88	A Modern Intro to Physics	465	401	111	32	16	1550	
...	...	...	...	...	..	..	...	
8184	Womens Law Journal	0	0	0	0	0	136	
8186	Wood Structures I	36	95	21	2	1	237	
8187	Wood Working	144	15	6	1	3	211	
8190	Workshop - Public Economics	84	0	0	0	0	166	
8191	Workshop in Dance Activity	1490	67	20	5	12	1862	
8192	Workshop in Econometrics	162	0	0	0	0	213	
8193	Workshop in Economic Theory	135	0	0	0	0	153	
8194	Workshop in Kinesiology	360	82	8	0	0	599	
8195	Workshop in Labor Economics	172	0	0	0	0	338	
8196	Workshop in Public Affairs	258	4	0	1	0	306	

8198	Workshop-Int'l Public Affairs	95	3	0	0	0	154
8199	Workshop-Physical Activity	403	30	3	1	3	612
8200	Workshop-Public Economics	113	0	0	0	0	191
8202	World Dance Cultures	253	48	6	3	0	401
8204	World Hunger & Malnutrition	744	313	87	41	12	1864
8206	World Regions in Global Contxt	368	543	99	45	62	1820
8208	World Vegetable Crops	134	84	48	15	7	430
8209	World/Postcolonial Lit-English	134	1	0	0	0	180
8213	Writing Travels	85	1	0	0	0	106
8217	Writing for TV & Film	58	75	1	1	0	244
8222	Writing, Rhetoric, & Literacy	170	23	3	0	4	230
8225	Wrkshp in Math Econ Theory	190	0	0	0	0	190
8226	Wrkshp-Industrl Organizatn	310	0	0	0	0	406
8227	Wrkshp-Internatl Economics	116	0	0	0	0	202
8228	Wrkshp-Schl Program Develop	174	1	0	0	0	188
8231	Yiddish Lit & Culture, America	74	17	0	1	2	171
8233	Yiddish Song and Jewish Exp	204	17	3	0	0	265
8234	Yng Adult Occs&Ther Interventn	106	32	2	0	0	277
8236	Young Adult Lit for Schools	103	3	0	0	0	114
8237	Young Adult Literature	184	4	2	0	0	213

	A rate	DF rate
7	0.582781	0.023841
8	0.596869	0.021526
9	0.401099	0.040293
10	0.431319	0.013736
11	0.262857	0.034286
23	0.639640	0.009009
24	0.358191	0.069682
25	0.449621	0.037920
26	0.700935	0.018692
27	0.398589	0.010582
28	0.883162	0.006873
29	0.427293	0.020134
32	0.671739	0.002174
34	0.933333	0.000000
39	0.220000	0.024000
41	0.755102	0.000000
48	0.331263	0.060041
49	0.507874	0.021654
52	0.576389	0.027778
60	0.424855	0.037572
62	0.253731	0.000649
63	0.272382	0.000000
64	0.490099	0.009901
65	0.311628	0.000000
72	0.384000	0.016000
74	0.601562	0.003906

```
79    0.384810  0.010127
86    0.390588  0.023529
87    0.279401  0.013542
88    0.300000  0.030968
...        ...       ...
8184  0.000000  0.000000
8186  0.151899  0.012658
8187  0.682464  0.018957
8190  0.506024  0.000000
8191  0.800215  0.009130
8192  0.760563  0.000000
8193  0.882353  0.000000
8194  0.601002  0.000000
8195  0.508876  0.000000
8196  0.843137  0.003268
8198  0.616883  0.000000
8199  0.658497  0.006536
8200  0.591623  0.000000
8202  0.630923  0.007481
8204  0.399142  0.028433
8206  0.202198  0.058791
8208  0.311628  0.051163
8209  0.744444  0.000000
8213  0.801887  0.000000
8217  0.237705  0.004098
8222  0.739130  0.017391
8225  1.000000  0.000000
8226  0.763547  0.000000
8227  0.574257  0.000000
8228  0.925532  0.000000
8231  0.432749  0.017544
8233  0.769811  0.000000
8234  0.382671  0.000000
8236  0.903509  0.000000
8237  0.863850  0.000000

[3765 rows x 9 columns]
```

We drop columns that are not useful to determine mental health issues.

```
In [190]: #dropped unnecessary columns
          df_survey=df_survey.drop(['Timestamp','phys_health_consequence','phys_health_intervie

In [191]: #extracted only people in CA and extracting the ones who do work in tech companies
          df_ca = df_survey[df_survey['state'] == 'CA']
          ca_tech = df_ca[df_ca['tech_company'] == 'Yes']

In [192]: #dropped unnecessary columns that doesn't need to be part of our analysis

          df_grad=df_grad.drop(['Major_code','Grad_median','Grad_P25','Grad_P75','Nongrad_media
```

```
In [193]: #drop rows with NaN
          capedf_nonstem = capedf_nonstem.dropna()
          #drop rows with NaN
          capedf = capedf.dropna()
```

10 Data Analysis & Results

We want to know how a higher number of study hours per week would affect students.

To do this for UCSD students, we got data from the CAPE (Course and Professor Evaluations) results websites and chose a variety of core classes for different majors.

We can look at classes with average hours each student studies and average grade each student recieved.

```
In [194]: capedf
```

Out[194]:

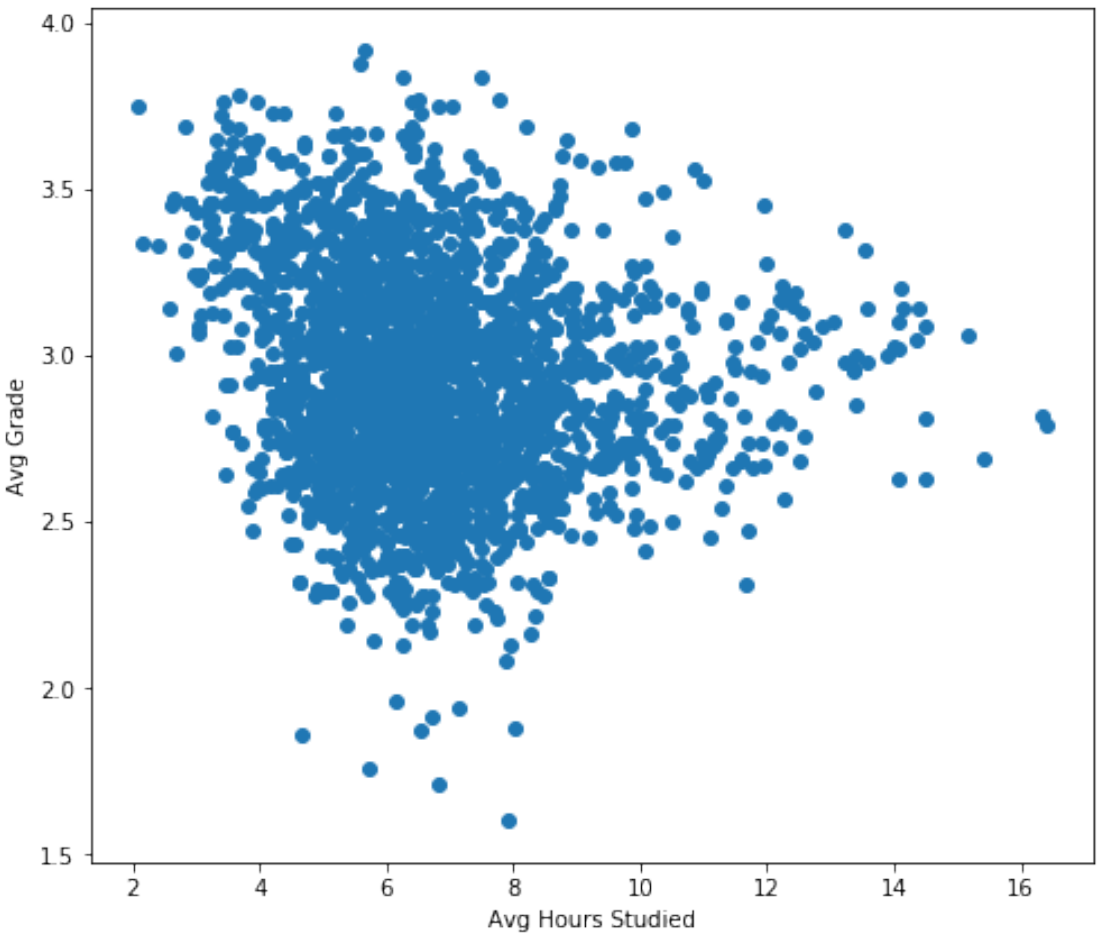
	Class	Avg Hours Studied	Avg Grade
2	BIBC103	5.93	3.31
3	BIBC103	6.60	3.18
4	BIBC103	5.58	2.91
5	BIBC103	9.44	3.15
6	BIBC103	8.96	3.17
7	BIBC103	4.50	2.90
8	BIBC103	5.62	3.12
9	BIBC103	6.32	3.46
10	BIBC103	6.50	3.36
11	BIBC103	9.59	3.01
12	BIBC103	6.23	3.40
13	BIBC103	6.83	3.24
14	BIBC103	5.36	2.96
15	BIBC103	6.20	3.13
16	BIBC103	6.38	3.12
18	BIBC103	9.58	2.88
19	BIBC103	5.23	3.03
20	BIBC103	5.77	3.46
21	BIBC103	7.68	3.16
22	BIBC103	6.03	2.85
23	BIBC103	6.79	3.22
24	BIBC103	7.10	3.23
25	BIBC103	6.97	3.17
26	BIBC103	6.34	3.19
27	BIBC103	8.74	3.14
29	BIBC103	9.86	3.27
30	BIBC103	7.45	2.52
31	BIBC103	6.36	3.21
32	BIBC103	8.88	3.19
33	BIBC103	8.27	3.22
...	...	...	...
2649	PHYS1A	4.50	3.46

```
2650   PHYS1A              3.61       3.46
2651   PHYS1A              2.90       3.46
2652   PHYS1A              3.64       3.46
2653   PHYS1A              3.25       3.46
2654   PHYS1A              3.50       3.46
2655   PHYS1A              4.10       3.46
2656   PHYS1A              4.50       3.46
2657   PHYS1A              2.90       3.46
2658   PHYS1A              3.50       3.46
2659   PHYS1A              6.42       2.53
2660   PHYS1A              6.73       2.86
2661   PHYS1A              6.49       2.86
2662   PHYS1A              6.37       2.86
2665   PHYS1A              5.91       2.77
2666   PHYS1A              5.44       3.47
2667   PHYS1A              3.98       2.94
2668   PHYS1A              4.66       2.94
2669   PHYS1A              4.13       2.94
2678   PHYS1A              5.52       2.46
2679   PHYS1A              5.32       2.87
2680   PHYS1A              5.13       2.91
2681   PHYS1A              4.65       2.91
2682   PHYS1A              5.39       2.91
2685   PHYS1A              4.82       3.22
2686   PHYS1A              5.04       2.88
2687   PHYS1A              4.63       2.88
2688   PHYS1A              5.05       2.93
2689   PHYS1A              4.59       2.93
2690   PHYS1A              4.25       2.93

[2440 rows x 3 columns]
```

Only looking at rows and columns are not enough for us to form causations, so we made a scatter graph. From the graph, we could tell that students studying higher number of hours per week does not mean that they would get a higher grade proportional to their efforts. Some students only studied less than 4 hours per week but got a pretty good grade whereas some students studies more than 10 hours per week but got a grade less than 3.0.

```
In [195]: plt.scatter(capedf['Avg Hours Studied'], capedf['Avg Grade'])
          plt.xlabel('Avg Hours Studied'); plt.ylabel('Avg Grade');
```

```
fig_size = plt.rcParams["figure.figsize"]
fig_size[0] = 8
fig_size[1] = 7
plt.rcParams["figure.figsize"] = fig_size
```



We want to look at the different classes by subject, so we create dataframes for each of them.

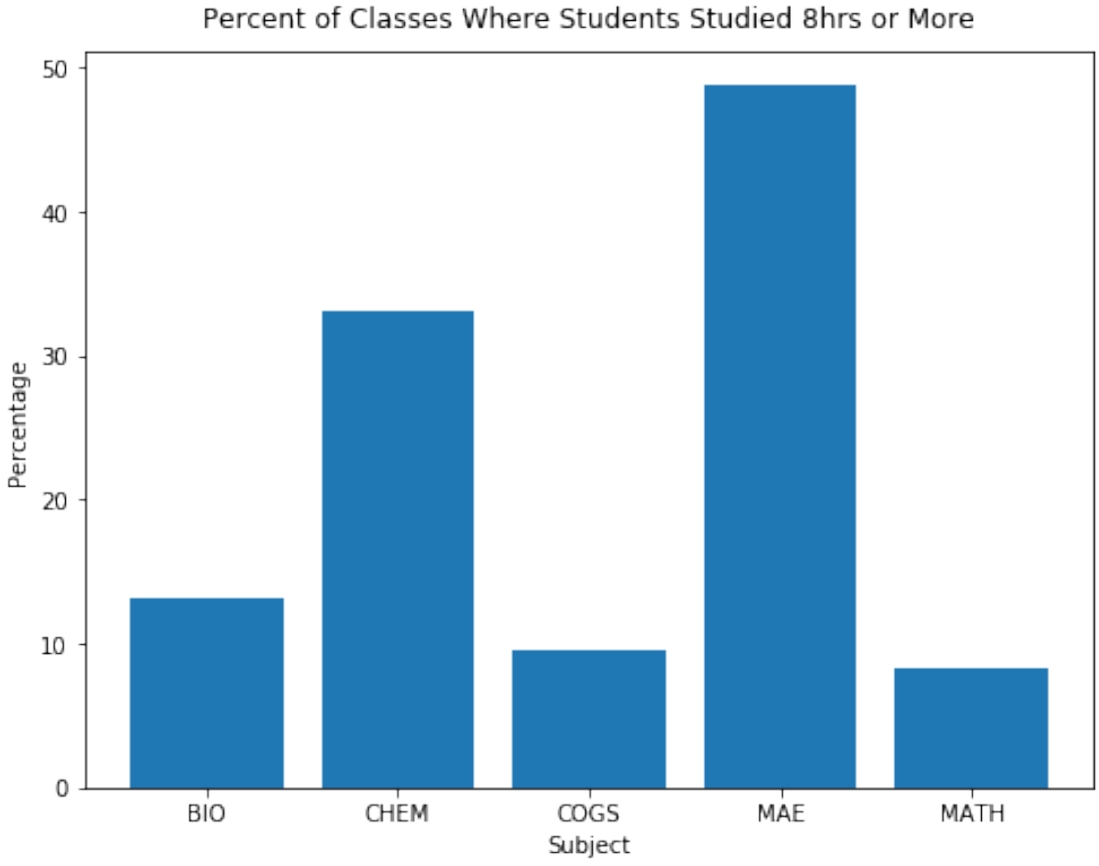
```
In [247]: df_bio = capedf[capedf['Class'].str.match('BI')]
          df_chem = capedf[capedf['Class'].str.match('CHEM')]
          df_cogs = capedf[capedf['Class'].str.match('COGS')]
          df_mae = capedf[capedf['Class'].str.match('MAE')]
          df_math = capedf[capedf['Class'].str.match('MATH')]
```

We can now look at how many classes reported studying for 8 hours or more on average per week based on their CAPE responses, by subject.

```
In [248]: bio_8hrs = df_bio[df_bio['Avg Hours Studied'] >= 8]
          chem_8hrs = df_chem[df_chem['Avg Hours Studied'] >= 8]
          cogs_8hrs = df_cogs[df_cogs['Avg Hours Studied'] >= 8]
          mae_8hrs = df_mae[df_mae['Avg Hours Studied'] >= 8]
          math_8hrs = df_math[df_math['Avg Hours Studied'] >= 8]

          bio_8hrs_p = len(bio_8hrs) / len(df_bio) * 100
          chem_8hrs_p = len(chem_8hrs) / len(df_chem) * 100
```

```
cogs_8hrs_p = len(cogs_8hrs) / len(df_cogs) * 100
mae_8hrs_p = len(mae_8hrs) / len(df_mae) * 100
math_8hrs_p = len(math_8hrs) / len(df_math) * 100

stem_subjects = ['BIO', 'CHEM', 'COGS', 'MAE', 'MATH']
stem_8hrs = [bio_8hrs_p, chem_8hrs_p, cogs_8hrs_p, mae_8hrs_p, math_8hrs_p]
```

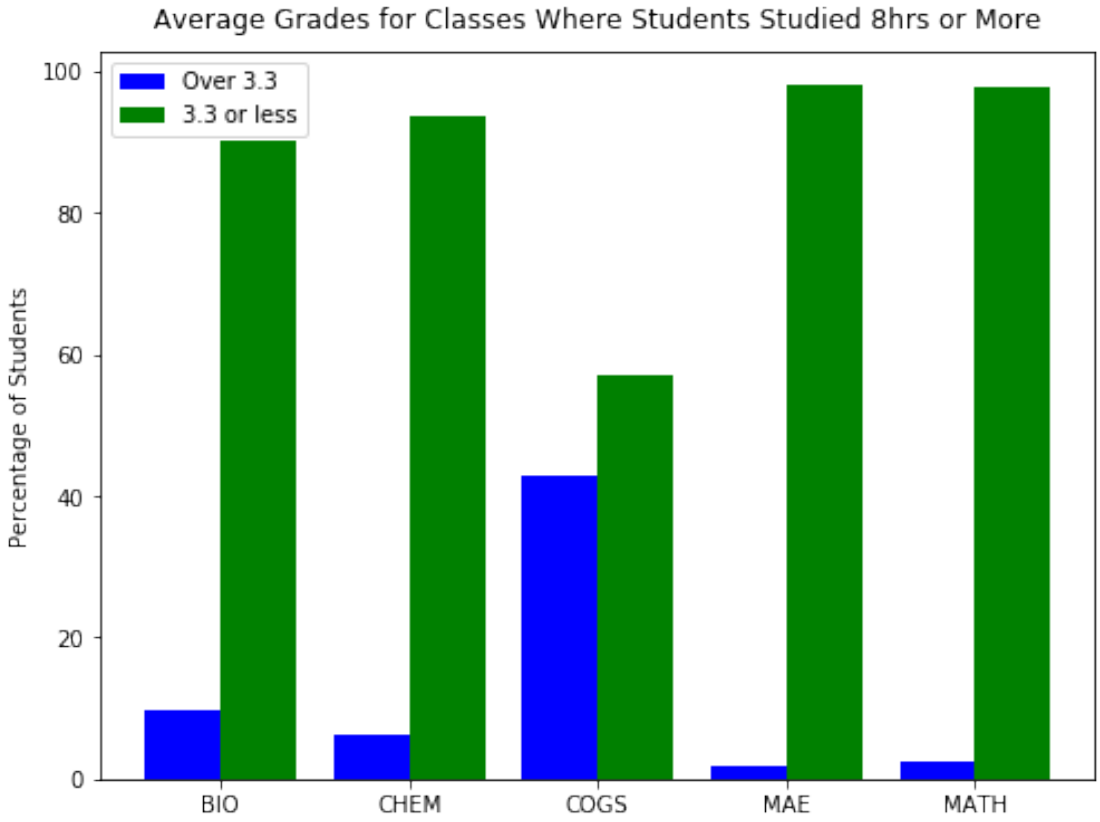
In [249]:
```
plt.bar(np.arange(len(stem_subjects)), stem_8hrs, align='center')
plt.xticks(ticks=np.arange(len(stem_subjects)), labels=stem_subjects)
plt.ylabel('Percentage'); plt.xlabel('Subject')
plt.title('Percent of Classes Where Students Studied 8hrs or More', pad=10)

fig_size = plt.rcParams["figure.figsize"]
fig_size[0] = 6
fig_size[1] = 4
plt.rcParams["figure.figsize"] = fig_size

plt.show()
```



From here, we can look at all the classes where students studied 8 hours or more per week on average, and out of those classes, how many had an average grade of at least a B, which at UCSD,

would be over a 3.3.

```
In [252]: bio_8hrs_h = len(df_bio[(df_bio['Avg Grade'] > 3.3) & (df_bio['Avg Hours Studied'] >=
          chem_8hrs_h = len(df_chem[(df_chem['Avg Grade'] > 3.3) & (df_chem['Avg Hours Studied
          cogs_8hrs_h = len(df_cogs[(df_cogs['Avg Grade'] > 3.3) & (df_cogs['Avg Hours Studied
          mae_8hrs_h = len(df_mae[(df_mae['Avg Grade'] > 3.3) & (df_mae['Avg Hours Studied'] >=
          math_8hrs_h = len(df_math[(df_math['Avg Grade'] > 3.3) & (df_math['Avg Hours Studied

          stem_grades_h = [bio_8hrs_h, chem_8hrs_h, cogs_8hrs_h, mae_8hrs_h, math_8hrs_h]
          stem_grades_lh =[]
          for i in range(len(stem_grades_h)):
              stem_grades_lh.append(100 - stem_grades_h[i])

In [254]: fig, ax = plt.subplots()
          indexes = np.arange(len(stem_subjects))
          bar_width = 0.4

          res1 = plt.bar(indexes, stem_grades_h, bar_width, align='center', color='blue', label
          res2 = plt.bar(indexes + bar_width, stem_grades_lh, bar_width, align='center', color=

          fig_size = plt.rcParams["figure.figsize"]
          fig_size[0] = 8
          fig_size[1] = 6
          plt.rcParams["figure.figsize"] = fig_size

          plt.xticks(ticks=indexes+(bar_width/2), labels=stem_subjects)
          plt.ylabel('Percentage of Students');
          plt.title('Average Grades for Classes Where Students Studied 8hrs or More', pad=10)
          plt.legend()

          plt.show()
```



In classes like MAE (Mechanical and Aerospace Engineering) and Math, over 90% of students who study 8 hours or more a week do not receive more than a 3.3. Students who are putting in the time and the effort to do well are not necessarily scoring high.

To get a better sense of how much STEM students study, we can compare their data with data from non-STEM students.

```
In [201]: capedf_nonstem
```

Out[201]:

	Class	Avg Hours Studied	Avg Grade
0	CGS100	4.14	3.56
3	CGS100	3.61	3.62
4	CGS100	5.58	3.08
5	CGS100	4.62	3.59
6	CGS100	7.50	3.66
7	CGS100	6.30	3.46
8	CGS100	6.57	3.69
9	CGS100	6.98	3.58
11	CGS100	5.50	3.45
13	CGS100	4.63	3.51
17	CGS101	5.58	3.78
18	CGS101	3.91	3.69
19	CGS101	5.42	3.83

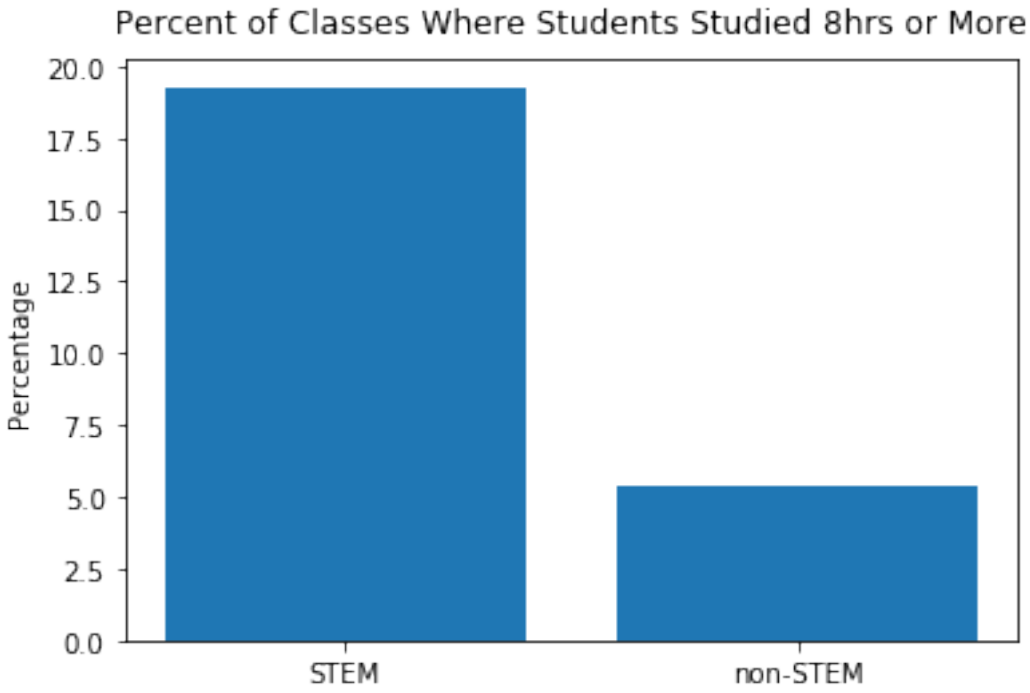
20	CGS101	4.27	3.48
21	CGS101	6.07	2.97
22	CGS101	5.03	3.39
23	CGS101	6.79	3.56
25	CGS101	6.20	3.39
27	CGS101	4.35	3.74
29	CGS101	5.20	3.74
30	HIEA13_	5.67	2.60
33	HIEA13_	5.33	3.62
34	HIEA13_	5.83	3.36
35	HIEA13_	7.20	3.71
36	HIEA13_	4.82	3.14
37	HIEA13_	8.39	3.86
38	HIEA13_	4.64	3.07
39	HIEA13_	5.17	2.94
40	HIEA13_	7.62	3.32
41	HIEA13_	8.94	2.94
..	...	...	...
782	SOCI104	6.50	3.54
783	SOCI104	3.50	3.35
785	SOCI104	6.25	3.93
786	SOCI104	3.83	3.24
787	SOCI104	5.75	3.60
789	SOCI104	7.61	3.56
791	SOCI104	7.50	3.50
792	SOCI104	7.36	3.18
793	SOCI104	7.93	3.40
794	SOCI104	4.90	3.62
795	SOCI104	5.13	3.66
796	SOCI104	8.83	3.42
798	SOCI104	5.88	3.60
799	SOCI104	6.50	2.71
800	SOCI104	6.50	3.41
803	SOCI104	4.72	3.88
806	SOCI104	7.27	3.29
807	SOCI104	4.13	3.51
808	SOCI104	3.59	3.77
814	SOCI106	5.10	3.02
816	SOCI106	5.30	3.00
817	SOCI106	4.50	3.11
818	SOCI106	7.10	3.79
819	SOCI106	8.21	3.82
820	SOCI106	3.70	3.16
823	SOCI106	3.05	3.44
824	SOCI106	3.83	3.30
826	VIS100	4.50	3.82
828	VIS100	5.30	4.00
836	VIS112	3.17	3.86

```
[616 rows x 3 columns]
```

We can make a graph to see how what percentage of classes had students studying 8 hours or more per week for both STEM and non-STEM clases.

```
In [202]: stem_8hrs = capedf[capedf['Avg Hours Studied'] >= 8]
          nonstem_8hrs = capedf_nonstem[capedf_nonstem['Avg Hours Studied'] >= 8]

          stem_8hrs_p = len(stem_8hrs) / len(capedf) * 100
          nonstem_8hrs_p = len(nonstem_8hrs) / len(capedf_nonstem) * 100
```

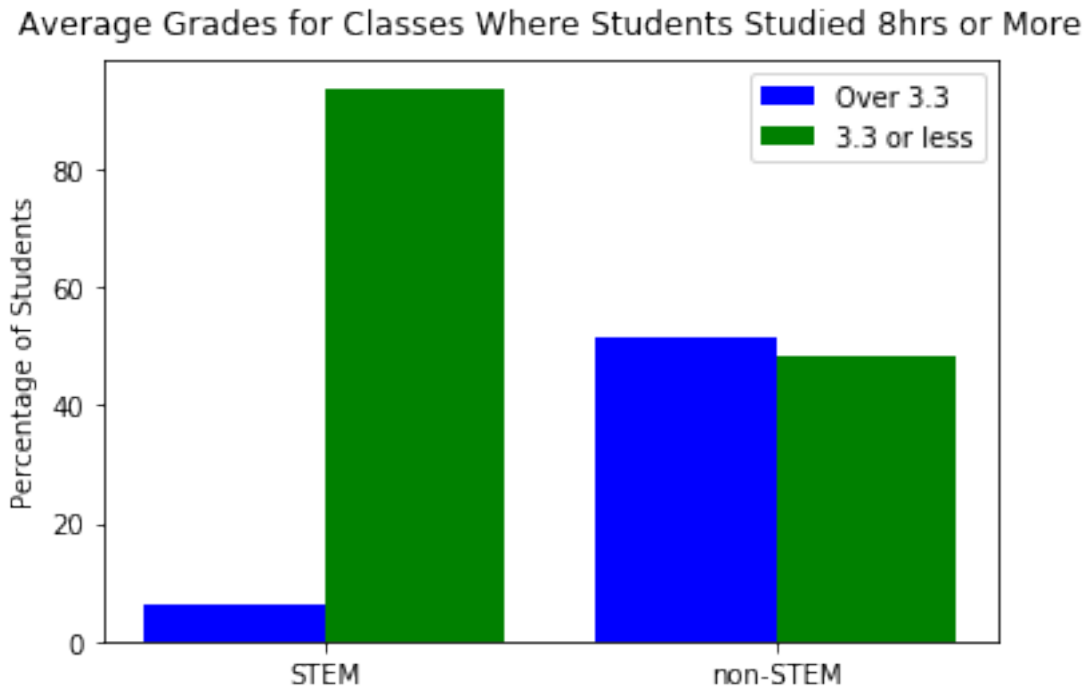
```
In [203]: plt.bar(np.arange(2), [stem_8hrs_p, nonstem_8hrs_p], align='center')
          plt.xticks(ticks=np.arange(2), labels=['STEM', 'non-STEM'])
          plt.ylabel('Percentage');
          plt.title('Percent of Classes Where Students Studied 8hrs or More', pad=10)

          plt.show()
```



It is clear from this graph that non-STEM students generally do not have to study as much, since only about 5% of classes reported an average of studying 8 hours or more per week.

Once again, we can look at all the classes where students studied 8 hours or more per week on average, and out of those classes, how many had an average grade over a 3.3, and we can compare STEM and non-STEM classes for these values.

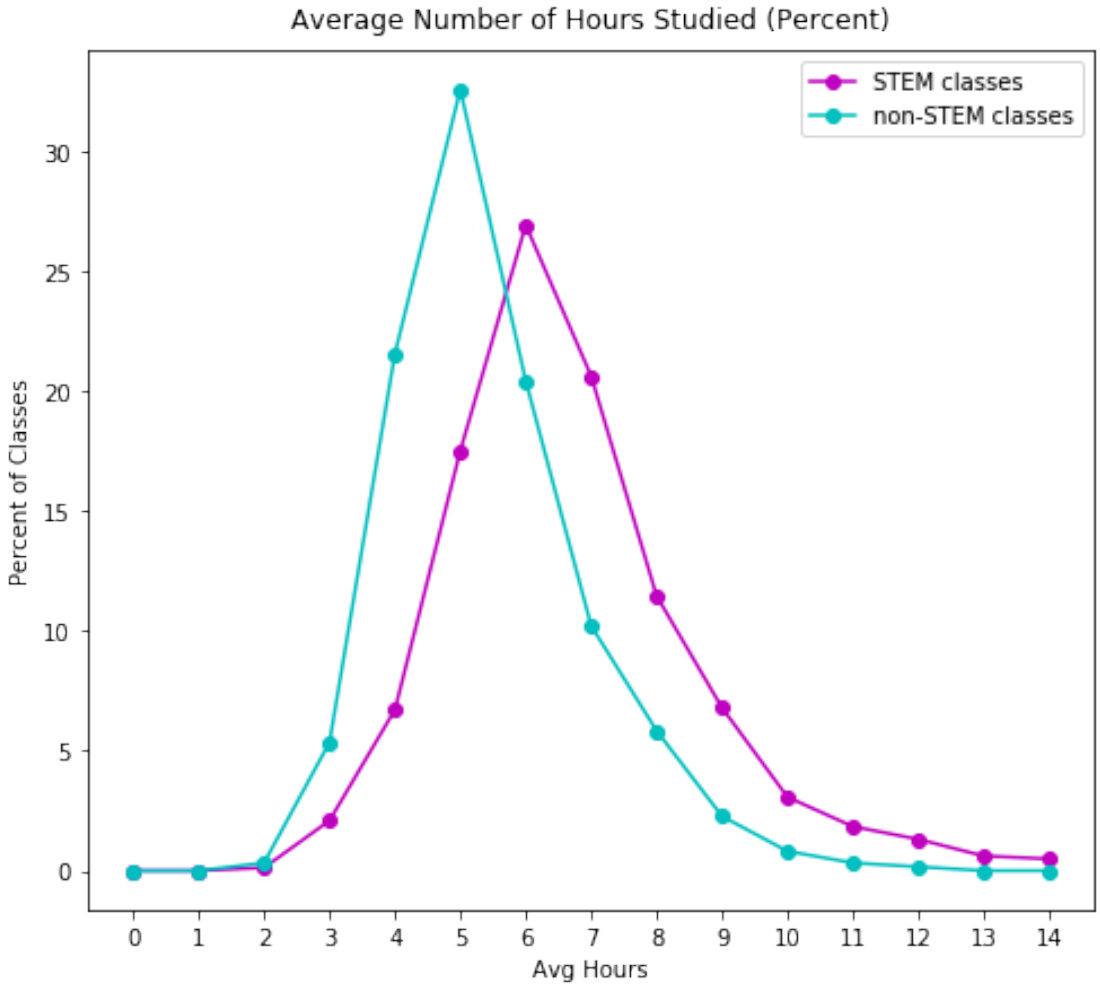
```
In [204]: stem_grades_p = len(stem_8hrs[stem_8hrs['Avg Grade'] > 3.3]) / len(stem_8hrs) * 100
          nonstem_grades_p = len(nonstem_8hrs[nonstem_8hrs['Avg Grade'] > 3.3]) / len(nonstem_8

          grades_h = [stem_grades_p, nonstem_grades_p]
          grades_lh = []
          for i in range(len(grades_h)):
              grades_lh.append(100 - grades_h[i])

In [205]: indexes = np.arange(2)
          bar_width = 0.4

          res1 = plt.bar(indexes, grades_h, bar_width, align='center', color='blue', label="Ove
          res2 = plt.bar(indexes + bar_width, grades_lh, bar_width, align='center', color='gree

          plt.xticks(ticks=indexes+(bar_width/2), labels=['STEM', 'non-STEM'])
          plt.ylabel('Percentage of Students');
          plt.title('Average Grades for Classes Where Students Studied 8hrs or More', pad=10)
          plt.legend()

          plt.show()
```



This graph shows that in general, STEM classes where students study for at least 8 hours a week do not usually receive average grades better than a B- (3.3), wheras non-STEM classes where students study at least 8 hours a week receive an average grade better than a B- about half the time.

Furthermore, we can look at how many hours students studied and what percentage of students studied for each specific amount of hours.

```
In [206]: stem_hrs = []
          nonstem_hrs = []

          bio_8hrs_h = len(df_bio[(df_bio['Avg Grade'] > 3.3) & (df_bio['Avg Hours Studied'] >=

          for i in range(0,15):
              stem_i_hrs = len(capedf[(capedf['Avg Hours Studied'] >= (i-0.5)) & (capedf['Avg H
              nonstem_i_hrs = len(capedf_nonstem[(capedf_nonstem['Avg Hours Studied'] >= (i-0.5

              stem_i_hrs_p = stem_i_hrs / len(capedf) * 100
              nonstem_i_hrs_p = nonstem_i_hrs / len(capedf_nonstem) * 100

              stem_hrs.append(stem_i_hrs_p)
              nonstem_hrs.append(nonstem_i_hrs_p)

In [237]: plt.plot([0,1,2,3,4,5,6,7,8,9,10,11,12,13,14],stem_hrs,'mo-', label='STEM classes')
          plt.plot([0,1,2,3,4,5,6,7,8,9,10,11,12,13,14],nonstem_hrs, 'co-', label='non-STEM cla

          plt.xticks(ticks=np.arange(15), labels=[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14])
          plt.title('Average Number of Hours Studied (Percent)', pad=10)
          plt.xlabel('Avg Hours'); plt.ylabel('Percent of Classes');
          plt.legend()

          fig_size = plt.rcParams["figure.figsize"]
          fig_size[0] = 8
          fig_size[1] = 7
          plt.rcParams["figure.figsize"] = fig_size

          plt.show()
```



In general, there is a trend that shows that STEM students spend more hours studying than non-STEM students do.

Based on what we have seen from these graphs, STEM students spend more time studying than non-STEM students do, but at the same time, STEM students still do not always receive good grades, and in fact, putting in over 8 hours a week still does not get a student a B most of the time. We would think that the more a student studies, the better they will do in the class, but according to these datasets, that is not the case. We see that STEM students study more, but we also see that at the same time, their grades are not as good compared to the grades of non-STEM students. Spending a lot of time studying is mentally taxing and can really take a toll on a person and their well-being, and it takes an even bigger toll when a student invests a lot of time and energy into a class, only to realize that they did not do as well as they maybe thought they could. The difficult and time-consuming nature of classes in the STEM field is not helping to improve the mental health of its students, especially seeing how many students study so much and still do not do well.

Now, moving on to a new dataset, to examine courses with the worst grades, we print out the highest D/F rate course and found out most of them are stem related.

```
In [208]: #Finding 100 highest D + F rate classes and store it
```

```
top_100_df = course_unique.nlargest(25, 'DF rate')
top_100_df
```

Out[208]:

	course	A	B	C	D	F	\
3543	Intermediate Algebra	265	277	222	129	98	
659	Animal Biology	2513	4450	4153	1802	908	
4206	Landscape Plants I	28	69	63	33	10	
5376	Political Sociology	57	25	14	14	11	
2130	Emergence of Human Culture	154	167	110	50	42	
2003	Economic Decision Analysis	44	67	51	34	11	
699	Appl Mathematical Analysis	270	316	332	150	72	
1023	Calc with Algebra & Trig II	641	760	774	329	147	
1028	Calculus&Analytic Geometry	3533	4054	3221	1306	925	
555	Algebra	1390	2187	1894	646	377	
7957	Trigonometry	472	560	480	230	107	
4347	Life of the Past	120	152	126	57	30	
3997	Intro:Prob&Markov Chain Mod	78	66	43	15	20	
368	Advanced Dynamics	81	122	76	40	18	
7086	Statics	958	1996	1218	484	330	
5758	Psychology of Perception	153	123	86	46	29	
5586	Principles-Wildlife Ecology	131	247	215	108	29	
2072	Elem Matrix&Linear Algebra	1096	989	621	319	167	
4612	Mechanical Vibrations	104	52	68	23	21	
2761	Fundamental Math Skills	53	52	42	16	11	
4585	Mathematical Logic	68	40	23	12	11	
6204	Satellite Dynamics	23	84	47	27	4	
1022	Calc with Algebra & Trig I	898	1157	939	358	192	
7246	Survey-Photogrphy:1839-1989	170	81	19	39	25	
5577	Principles-Food Preservation	40	120	98	33	5	

	Total students	A rate	DF rate
3543	1220	0.217213	0.186066
659	16653	0.150904	0.162733
4206	278	0.100719	0.154676
5376	167	0.341317	0.149701
2130	628	0.245223	0.146497
2003	314	0.140127	0.143312
699	1572	0.171756	0.141221
1023	3474	0.184514	0.137018
1028	16624	0.212524	0.134204
555	7700	0.180519	0.132857
7957	2626	0.179741	0.128332
4347	697	0.172166	0.124821
3997	286	0.272727	0.122378
368	478	0.169456	0.121339
7086	6764	0.141632	0.120343
5758	628	0.243631	0.119427
5586	1153	0.113617	0.118820

```
2072            4147  0.264287  0.117193
4612             376  0.276596  0.117021
2761             231  0.229437  0.116883
4585             197  0.345178  0.116751
6204             269  0.085502  0.115242
1022            4899  0.183303  0.112268
7246             571  0.297723  0.112084
5577             345  0.115942  0.110145
```

To make a more clear image we also calculate top 100 classes with most A grade and we found out that most of the highest A rate courses are literature/art/non stem related. even if they are stem related, they are intro (lower division) class, as we can see in the table below.

In [209]:
```
# find the highest A rate course
top_100_a = course_unique.nlargest(25, 'A')
top_100_a
```

Out[209]:

	course	A	B	C	D	F	\
6998	Special Topics	20647	2098	364	88	87	
4883	Music in Performance	16229	268	78	20	44	
1580	Contemporary Topics	13458	2124	277	40	32	
2854	General Physics	8951	10447	6678	816	258	
2841	General Chemistry	8010	9928	5704	1369	514	
3380	Independent Study	7904	112	11	1	12	
1848	Directed Study	6751	95	25	4	21	
3743	Intro to College Composition	6431	560	45	15	17	
8035	Varsity Band	6308	5	0	0	0	
4034	Introduction to Psychology	6301	4526	5281	1872	613	
5580	Principles-Microeconomics	5582	7230	3885	1091	348	
3283	Human Dev: Ed Effectiveness	5464	274	79	43	70	
8005	University Band	5302	46	9	3	6	
5671	Professional Communication	5099	222	12	2	10	
1669	Cult Anthro&Human Diversity	4844	4596	1023	225	102	
3977	Intro-Statistical Methods	4641	3267	1643	612	370	
4050	Introductory Biology	4559	10708	3923	915	219	
3984	Intro-Theatre & Dramatic Lit	4337	969	176	81	49	
563	All-Univ String Orchestra	4329	21	9	5	11	
1024	Calc--Functns of Variables	4320	4233	3002	1003	575	
3676	Intro Organic Chemistry	4051	5315	4744	1116	724	
2739	Freshman Composition	4004	1035	109	18	46	
3025	Health Care: Intrdis Appr	3758	106	17	18	14	
8183	Womens Bodies-Hlth&Disease	3716	895	178	38	40	
1028	Calculus&Analytic Geometry	3533	4054	3221	1306	925	

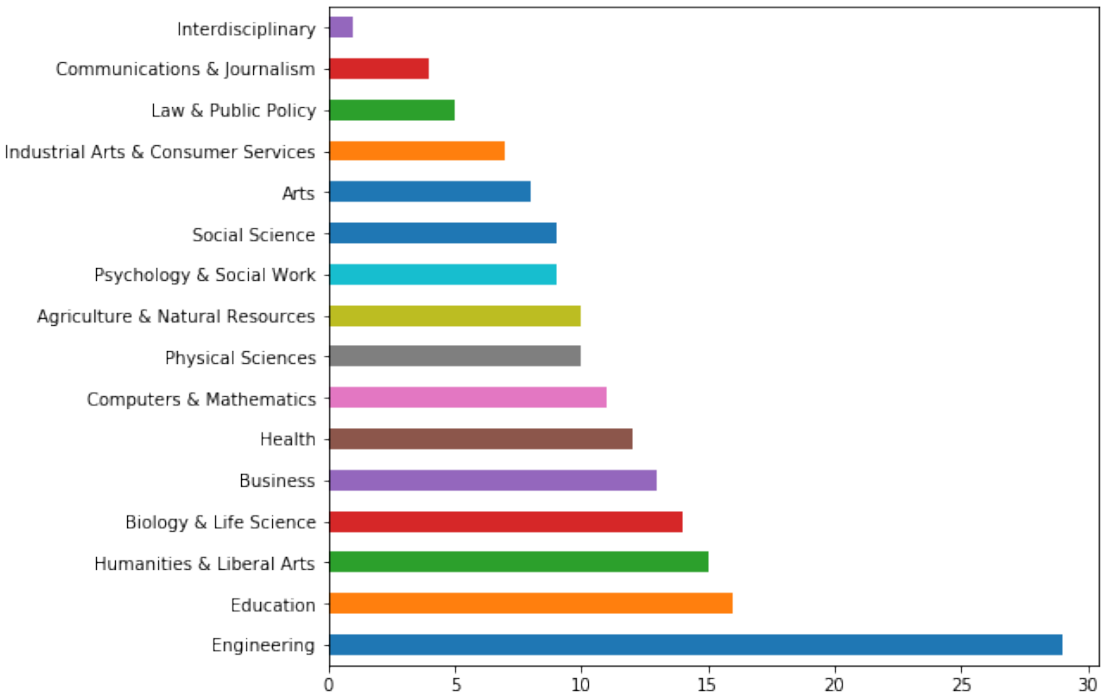
	Total students	A rate	DF rate
6998	35837	0.576136	0.004883
4883	18521	0.876249	0.003456
1580	21117	0.637306	0.003410

```
2854            44341  0.201867  0.024221
2841            32856  0.243791  0.057311
3380             8671  0.911544  0.001499
1848             7702  0.876526  0.003246
3743             9181  0.700468  0.003485
8035             6366  0.990889  0.000000
4034            29219  0.215647  0.085047
5580            29992  0.186116  0.047979
3283             6640  0.822892  0.017018
8005             5488  0.966108  0.001640
5671             8091  0.630206  0.001483
1669            16690  0.290234  0.019593
3977            14622  0.317398  0.067159
4050            25021  0.182207  0.045322
3984             8068  0.537556  0.016113
563              4484  0.965433  0.003568
1024            17479  0.247154  0.090280
3676            19623  0.206441  0.093768
2739             7837  0.510910  0.008166
3025             4363  0.861334  0.007334
8183             7268  0.511282  0.010732
1028            16624  0.212524  0.134204
```

Once again, we see that STEM students do not receive the best grades. Combining it with the fact that they also spend a lot of time studying, as we saw earlier, it further supports the fact that STEM students have to work really hard, and they do, sseing the hours they put in, but they are still not getting the best grades. It shows how challenging the STEM field is and how much students are being pushed to do more and try to do better, and that's really not always good, considering how STEM students still do not get the best grades even after studying a lot.

Now, let's look at different majors and their employment rate statistics.

```
In [210]: #major category count
          df_grad['Major_category'].value_counts()
          df_grad['Major_category'].value_counts().plot.barh()
          plt.show()
```



We dropped unnecessary columns that are not directly related to our analysis and only kept employement and unemployment for both graduates and nongraduates.

```
In [211]: # grouped by the major categories, without repetition.
          unique_majors=df_grad.groupby('Major_category', as_index=False).sum()

          unique_majors
```

Out[211]:

	Major_category	Grad_total	Grad_sample_size	\
0	Agriculture & Natural Resources	241342	4985	
1	Arts	580416	8410	
2	Biology & Life Science	1656556	33497	
3	Business	2718897	53120	
4	Communications & Journalism	462880	8639	
5	Computers & Mathematics	919817	17155	
6	Education	3945300	53944	
7	Engineering	2132524	42469	
8	Health	1468337	26089	
9	Humanities & Liberal Arts	2825975	47225	
10	Industrial Arts & Consumer Services	317219	5601	
11	Interdisciplinary	14405	318	
12	Law & Public Policy	280852	5448	
13	Physical Sciences	1052485	19360	
14	Psychology & Social Work	1630545	28044	
15	Social Science	1839710	35097	

	Grad_employed	Grad_unemployed	Nongrad_total	Nongrad_employed	\
0	179287	4995	599239	453541	
1	422450	24559	1657523	1194452	
2	1365336	32022	1145597	831399	
3	2124495	101994	9345634	7123852	
4	368390	17733	1635679	1285961	
5	716607	29062	1676169	1332370	
6	2437166	66938	4488291	2659824	
7	1634563	65073	3382085	2483802	
8	1148800	25962	2768323	2058011	
9	1986572	85033	3448921	2289696	
10	239338	8983	939696	680035	
11	12708	261	41018	32600	
12	224832	10011	831050	664417	
13	770365	24030	952098	656340	
14	1255928	49428	1795602	1271014	
15	1381570	60528	2439689	1720445	

	Nongrad_unemployed
0	16437
1	88900
2	44656
3	393222
4	86476
5	70960
6	111875
7	132162
8	63621
9	154239
10	33771
11	2573
12	36224
13	34404
14	87224
15	111390

We calculate the employment rates and unemployment rates for both graduates and nongraduates and combine them to get a overall employment rates.

In [212]:
```
#this is the unemployement and employement for both graduates and nongraduates
#created a new columns for the rates

unique_majors['employed_rates_g']=unique_majors['Grad_employed']/unique_majors['Grad_
unique_majors['unemployed_rates_g']=unique_majors['Grad_unemployed']/unique_majors['

unique_majors['employed_rates_non']=unique_majors['Nongrad_employed']/unique_majors[
unique_majors['unemployed_rates_non']=unique_majors['Nongrad_unemployed']/unique_maj
```

```
unique_majors
```

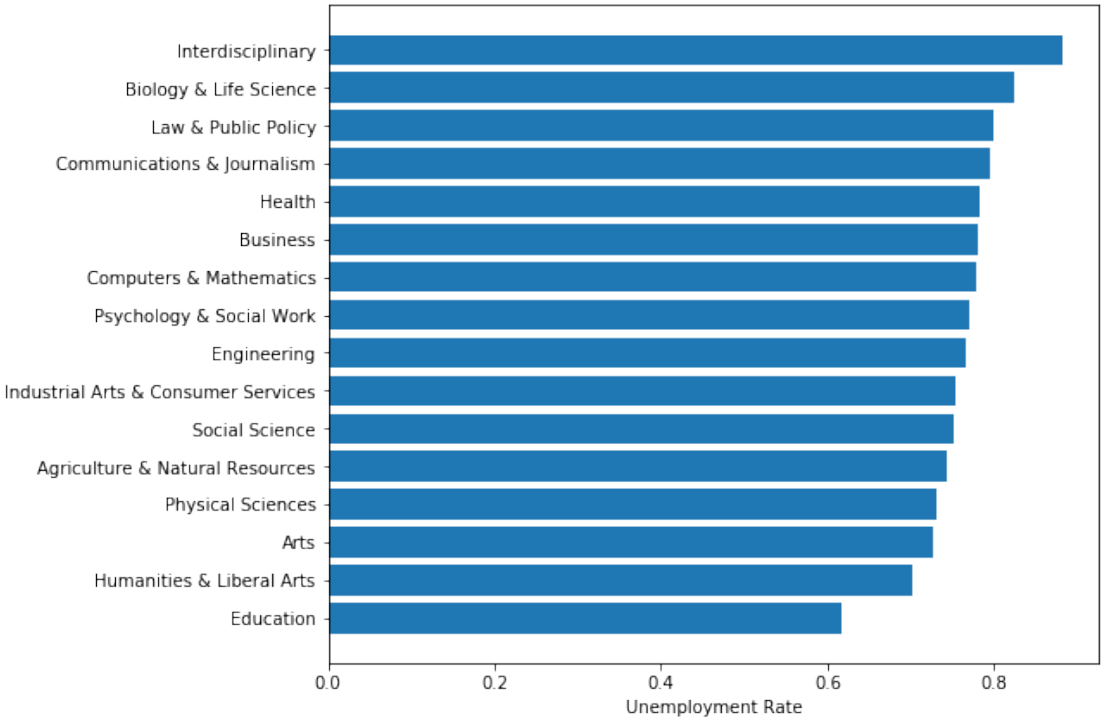
Out[212]:

	Major_category	Grad_total	Grad_sample_size	Grad_employed	Grad_unemployed	Nongrad_total	Nongrad_employed	Nongrad_unemployed	employed_rates_g	unemployed_rates_g
0	Agriculture & Natural Resources	241342	4985	179287	4995	599239	453541	16437	0.742875	0.020697
1	Arts	580416	8410	422450	24559	1657523	1194452	88900	0.727840	0.042313
2	Biology & Life Science	1656556	33497	1365336	32022	1145597	831399	44656	0.824202	0.019330
3	Business	2718897	53120	2124495	101994	9345634	7123852	393222	0.781381	0.037513
4	Communications & Journalism	462880	8639	368390	17733	1635679	1285961	86476	0.795865	0.038310
5	Computers & Mathematics	919817	17155	716607	29062	1676169	1332370	70960	0.779076	0.031595
6	Education	3945300	53944	2437166	66938	4488291	2659824	111875	0.617739	0.016967
7	Engineering	2132524	42469	1634563	65073	3382085	2483802			
8	Health	1468337	26089	1148800	25962	2768323	2058011			
9	Humanities & Liberal Arts	2825975	47225	1986572	85033	3448921	2289696			
10	Industrial Arts & Consumer Services	317219	5601	239338	8983	939696	680035			
11	Interdisciplinary	14405	318	12708	261	41018	32600			
12	Law & Public Policy	280852	5448	224832	10011	831050	664417			
13	Physical Sciences	1052485	19360	770365	24030	952098	656340			
14	Psychology & Social Work	1630545	28044	1255928	49428	1795602	1271014			
15	Social Science	1839710	35097	1381570	60528	2439689	1720445			

```
7                132162          0.766492          0.030515
8                 63621          0.782382          0.017681
9                154239          0.702969          0.030090
10                33771          0.754488          0.028318
11                 2573          0.882194          0.018119
12                36224          0.800536          0.035645
13                34404          0.731949          0.022832
14                87224          0.770250          0.030314
15               111390          0.750972          0.032901

    employed_rates_non  unemployed_rates_non
0             0.756862              0.027430
1             0.720625              0.053634
2             0.725734              0.038981
3             0.762265              0.042075
4             0.786194              0.052869
5             0.794890              0.042335
6             0.592614              0.024926
7             0.734400              0.039077
8             0.743414              0.022982
9             0.663888              0.044721
10            0.723676              0.035938
11            0.794773              0.062729
12            0.799491              0.043588
13            0.689362              0.036135
14            0.707848              0.048576
15            0.705190              0.045657
```

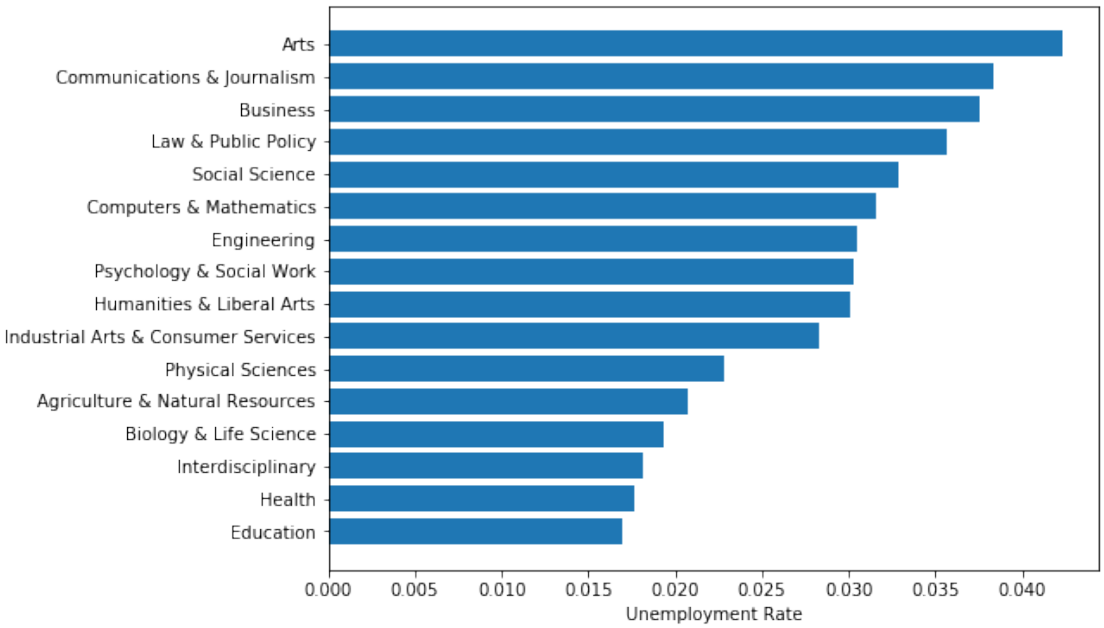
The graph below shows that employment rates are relativly high for STEM-related majors.

In [238]:
```
#bar graph for employed rates for graduates

unique_majors=unique_majors.sort_values(by=['employed_rates_g'])
plt.xlabel('Unemployment Rate');

fig_size = plt.rcParams["figure.figsize"]
fig_size[0] = 6
fig_size[1] = 4
plt.rcParams["figure.figsize"] = fig_size

plt.barh(unique_majors['Major_category'],unique_majors['employed_rates_g'])
plt.show()
```

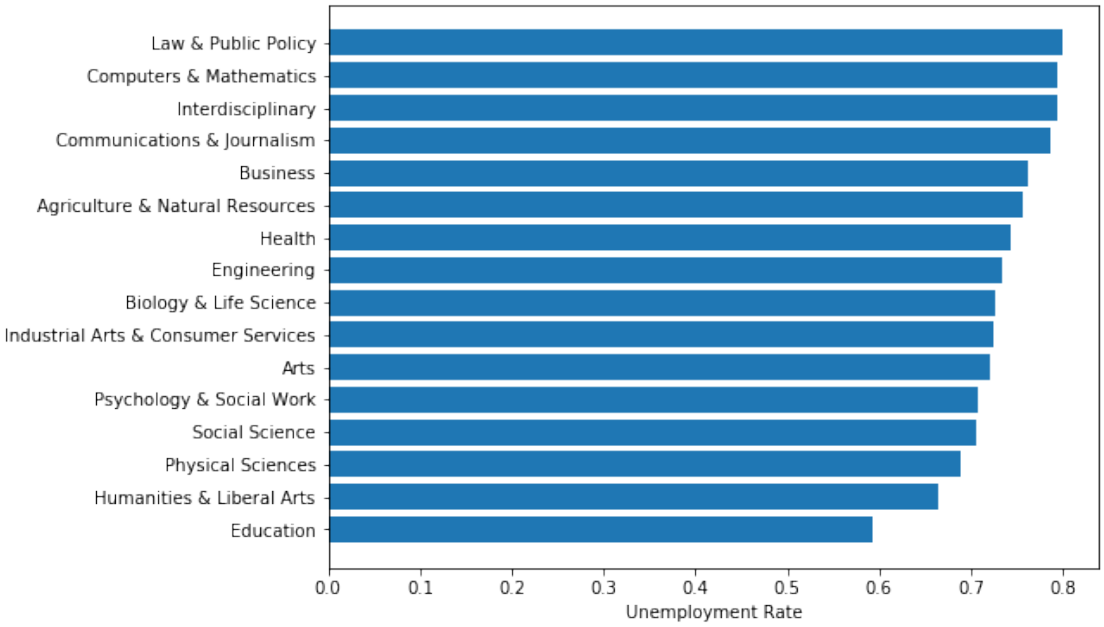


The graph below shows that unemployment rates are relatively high for non STEM-related majors.

```
In [244]: #bar graph for unemployement rates for graduates

          unique_majors=unique_majors.sort_values(by=['unemployed_rates_g'])
          plt.xlabel('Unemployment Rate');

          fig_size = plt.rcParams["figure.figsize"]
          fig_size[0] = 8
          fig_size[1] = 6
          plt.rcParams["figure.figsize"] = fig_size

          plt.barh(unique_majors['Major_category'],unique_majors['unemployed_rates_g'])
          plt.show()
```

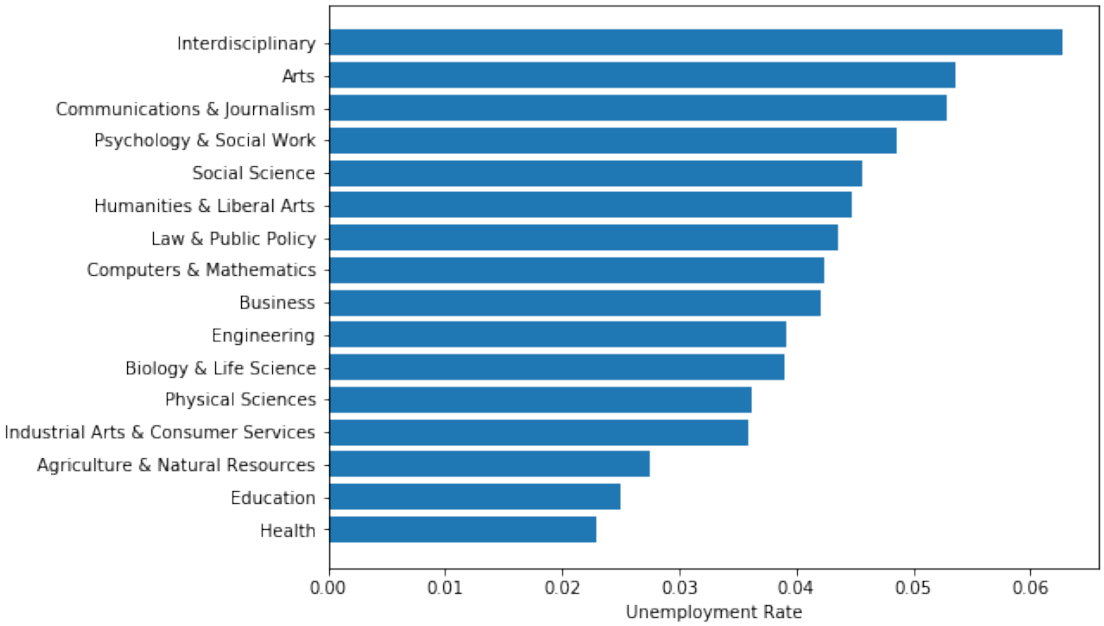


In [243]:
```
#bar graph for employed rates for nongraduates

unique_majors=unique_majors.sort_values(by=['employed_rates_non'])
plt.xlabel('Unemployment Rate');

plt.barh(unique_majors['Major_category'],unique_majors['employed_rates_non'])

plt.show()
```



STEM-related majors have relatively high employment rates, even with nongraduates, as seen below.

```
In [245]: #bar graph for unemployed rates for nongraduates

          unique_majors=unique_majors.sort_values(by=['unemployed_rates_non'])
          plt.xlabel('Unemployment Rate');

          plt.barh(unique_majors['Major_category'],unique_majors['unemployed_rates_non'])

          plt.show()
```

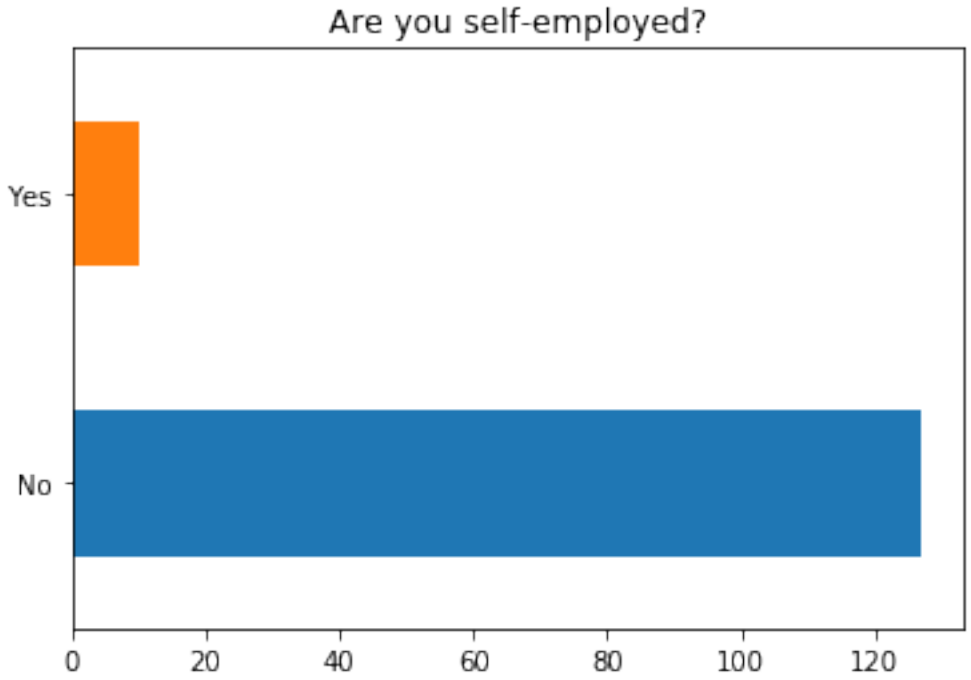


As we saw earlier, STEM students spend a lot of time studying but don't always necessarily do well, but looking at data for employment rates, it turns out that it actually does end up being worth it to some extent, since it gets them jobs after college, as we saw from this data. This is what they are working towards in college and this is why they work really hard and sometimes overexert themselves, and although it does usually work out for them because they can get good jobs, this mentality that they develop in college is something that they can also carry to the workplace. STEM majors are always striving to be very knowledgeable in their field and to be the best, because that's what will generally get them a better job and more money. This notion of constantly needing to be better and work harder is good in the sense that people will take their work seriously and always be trying to improve, but it's also bad because it is very possible to try and do more than your brain can handle. This is the kind of mindset that many college graduates bring to the workplace, which could be one reason for mental health conditions later on in the workplace.

Below are lists of bar charts of survey csv file for the mental health in tech kaggle page. We used value_counts method to obtain responses from each mental health related question in order to figure out the range of mental health issues in Tech insdustries.

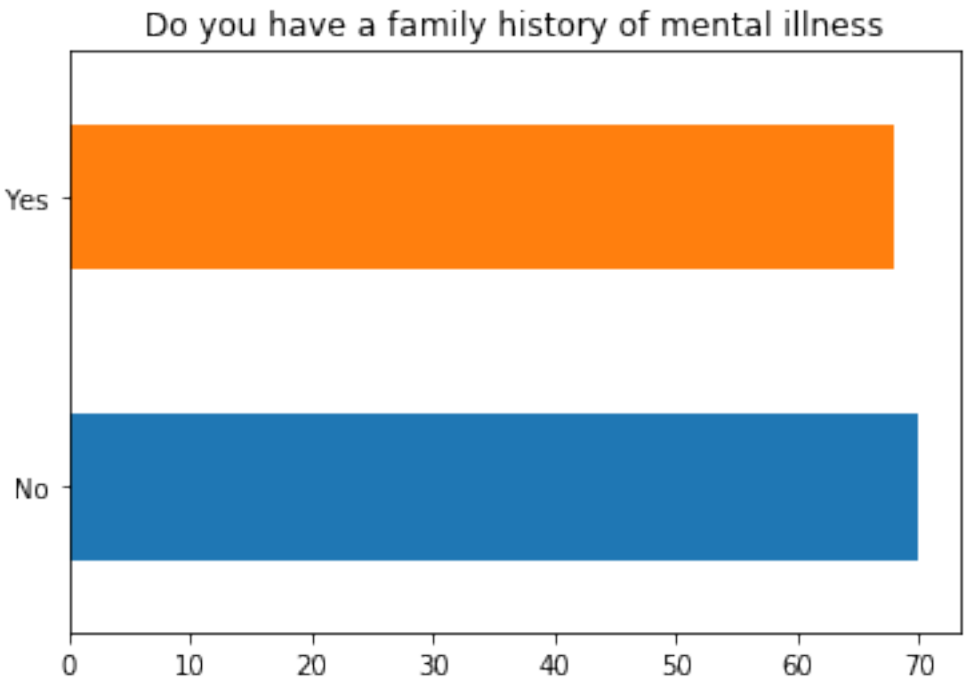
```
In [217]: #plotted each of the categorical responses
          #are they self-employed?
          df_ca['self_employed'].value_counts().plot(x="# of responders",y="responses",title='
          df_ca['self_employed'].value_counts()
```

```
Out[217]: No     127
          Yes     10
          Name: self_employed, dtype: int64
```

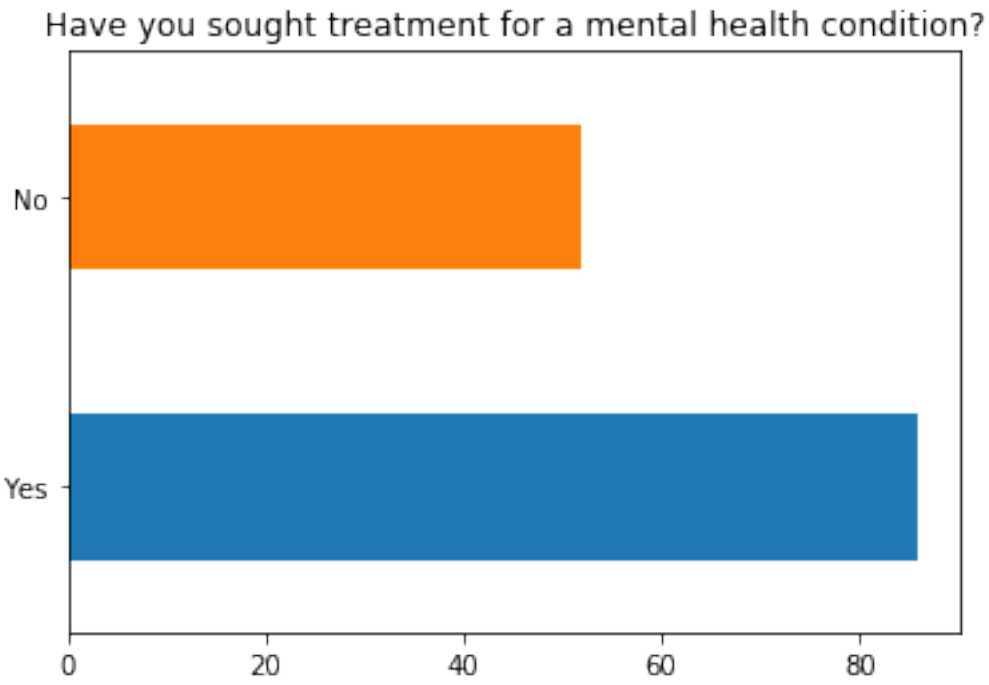



```
In [218]: #column = family history
          df_ca['family_history'].value_counts().plot(title='Do you have a family history of m
          df_ca['family_history'].value_counts()
```

```
Out[218]: No     70
          Yes    68
          Name: family_history, dtype: int64
```



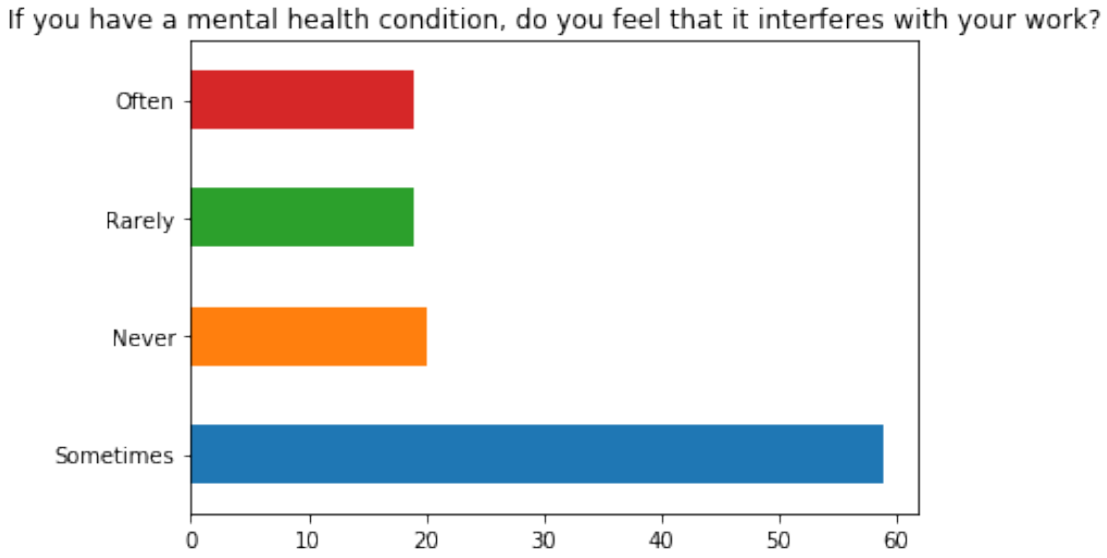

```
In [219]: #treatment
          df_ca['treatment'].value_counts().plot(title='Have you sought treatment for a mental
          df_ca['treatment'].value_counts()

Out[219]: Yes    86
          No     52
          Name: treatment, dtype: int64
```



As we can see, most of the people who answered these questions have real mental health conditions and have gone to a professional for treatment.

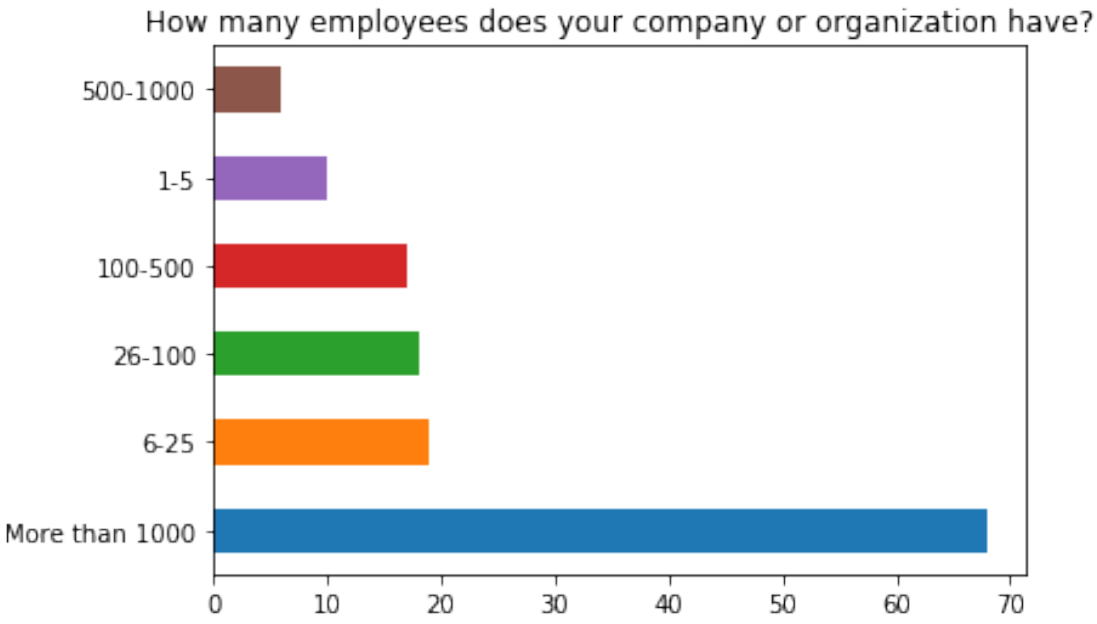
Below is a bar chart that asks employees do they feel that mental health condition interferes with their work if they had one. We assume that people have answered this question had a mental health conditon. The data of mental health in tech survey included surveys of 1260 people who was working in tech industries in 2015. Based on this data, it shows that around 10 percent of employers were experiencing mental health proplem in Tech industries.

```
In [220]: #interference
          df_ca['work_interfere'].value_counts().plot(title='If you have a mental health condi
          df_ca['work_interfere'].value_counts()

Out[220]: Sometimes    59
          Never        20
          Rarely       19
          Often        19
          Name: work_interfere, dtype: int64
```



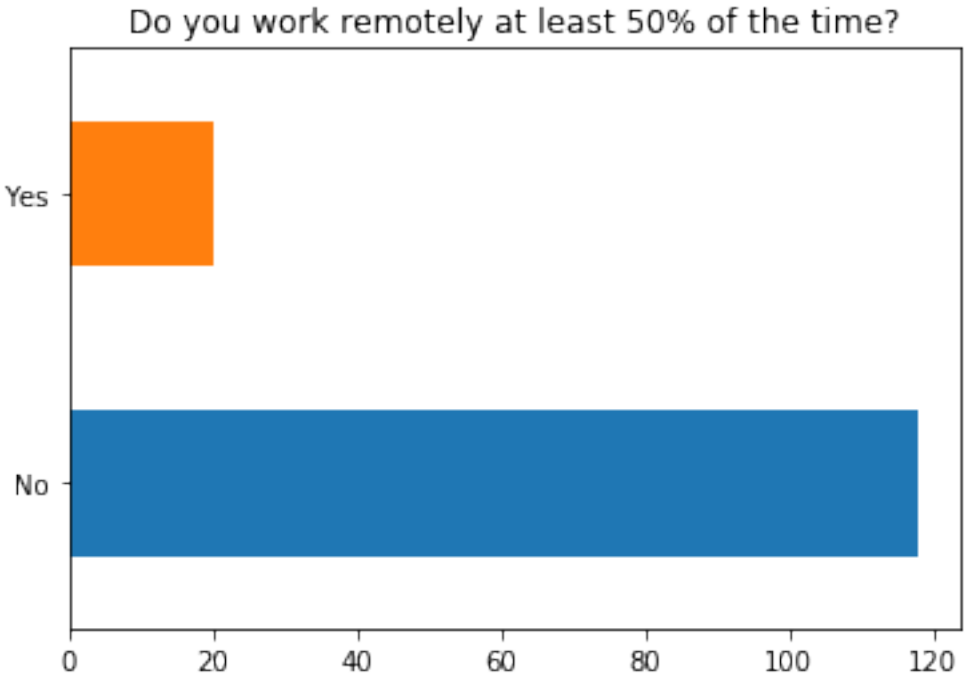
From the bar chart below, we could see that most of the companies have employees more than 1000

```
In [221]: #number of employees
          df_ca['no_employees'].value_counts().plot(title='How many employees does your company
          df_ca['no_employees'].value_counts()

Out[221]: More than 1000    68
          6-25              19
          26-100            18
          100-500           17
          1-5               10
          500-1000           6
          Name: no_employees, dtype: int64
```



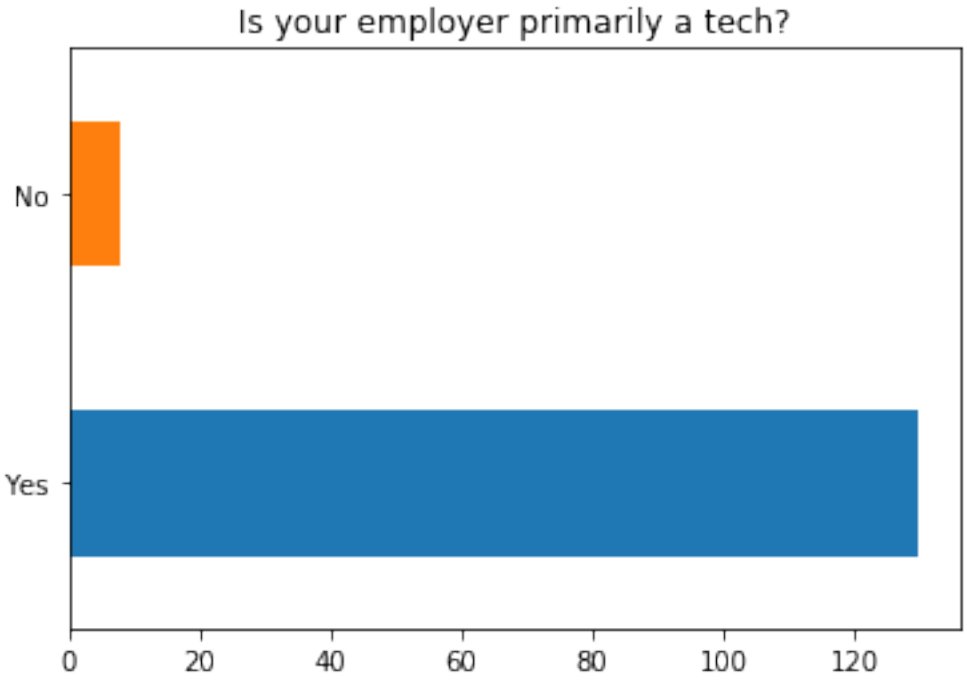

```
In [222]: #work interferance
          df_ca['remote_work'].value_counts().plot(title='Do you work remotely at least 50% of
          df_ca['remote_work'].value_counts()
```

```
Out[222]: No     118
          Yes     20
          Name: remote_work, dtype: int64
```



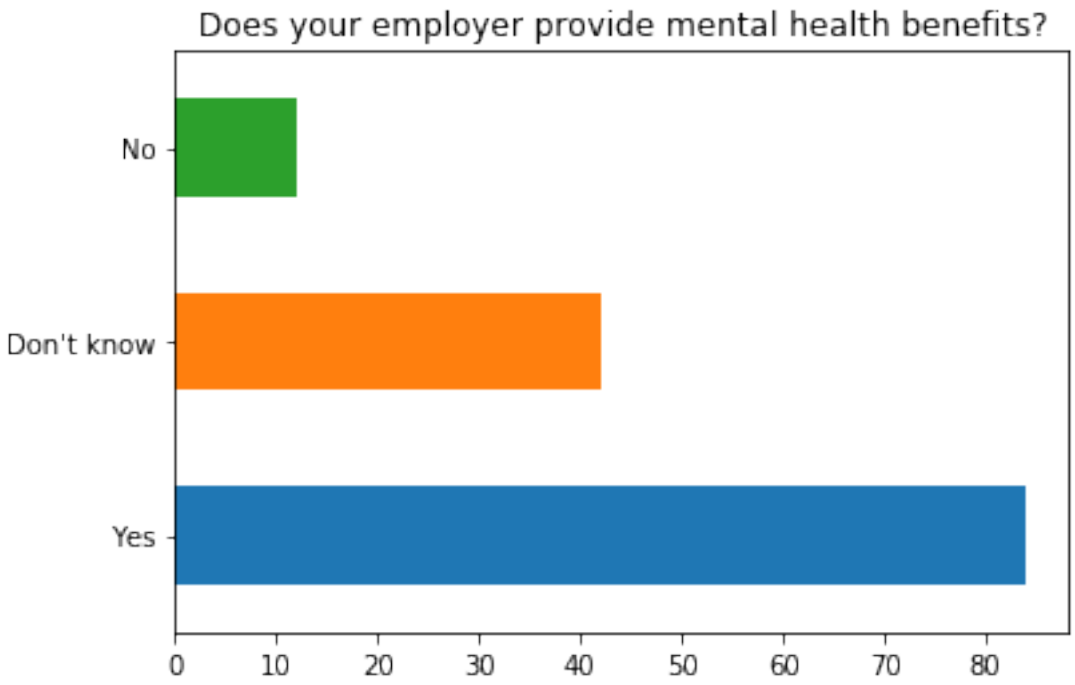
Most employees do not work remotely, meaning that they are in a work setting around their coworkers every day.

```
In [223]: #do they work for a tech company or not
          df_ca['tech_company'].value_counts().plot(title='Is your employer primarily a tech?'
          df_ca['tech_company'].value_counts()

Out[223]: Yes    130
          No       8
          Name: tech_company, dtype: int64
```



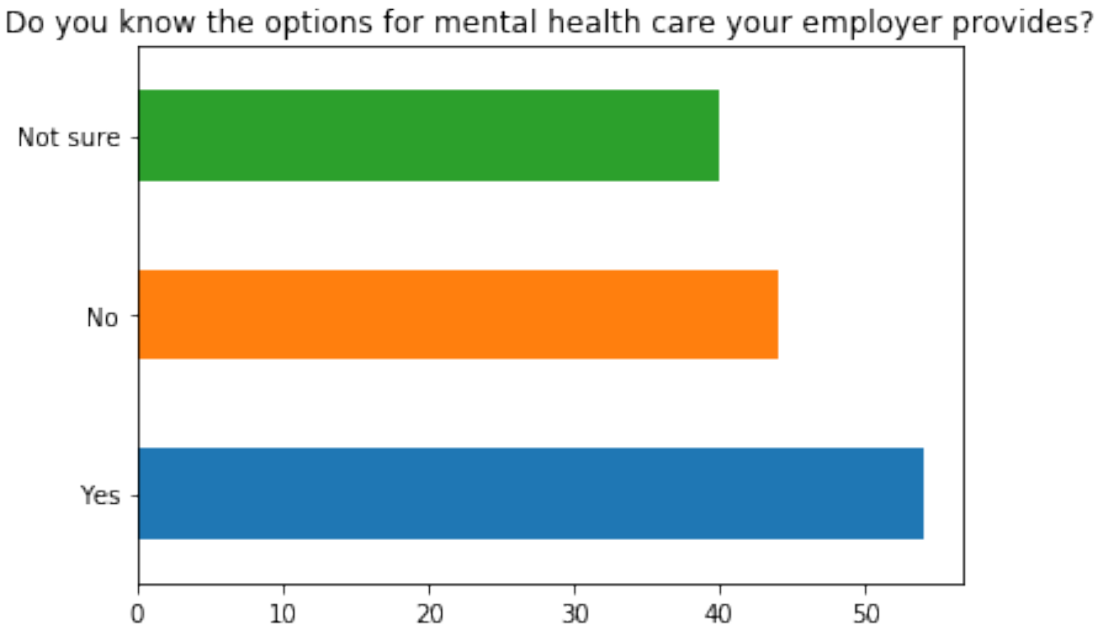
Most of the companies in tis dataset are tech companies.

```
In [224]: # benefits
          df_ca['benefits'].value_counts().plot(title='Does your employer provide mental healt
          df_ca['benefits'].value_counts()

Out[224]: Yes           84
          Don't know    42
          No            12
          Name: benefits, dtype: int64
```



This bar shows that most of companies do provide mental health benefits, which implies that mental health condition in tech industries are not rare, but rather many employers have realized its ponderance.

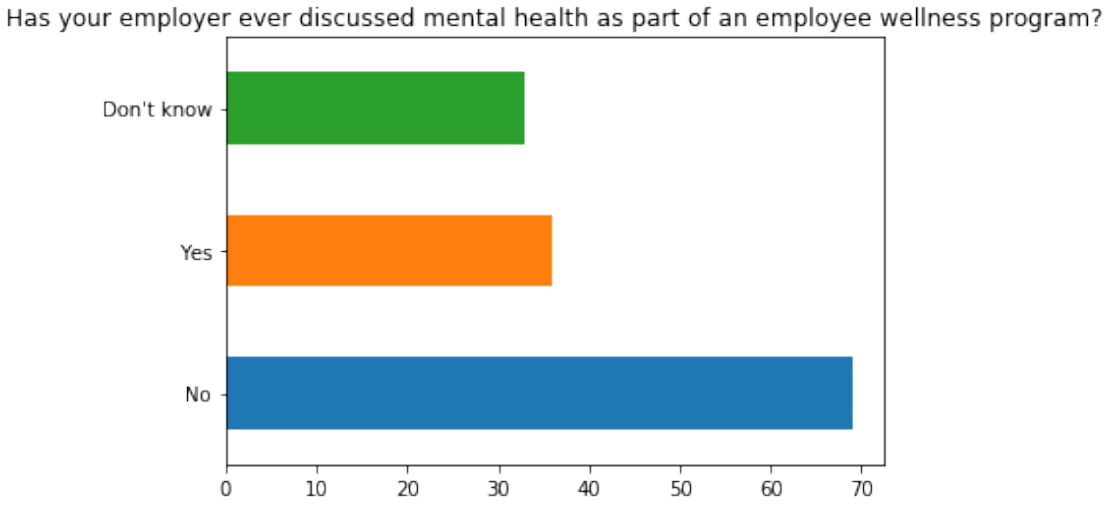
```
In [225]: #care options
          df_ca['care_options'].value_counts().plot(title='Do you know the options for mental h
          df_ca['care_options'].value_counts()

Out[225]: Yes         54
          No          44
          Not sure    40
          Name: care_options, dtype: int64
```



More than half the people who answered say that they do not know or not sure the options for mental health care their employer provides. We could assume that more than half of the employees are not aware of mental health issues, so thay either do not seek for help, or they do not know that they have mental health issues.

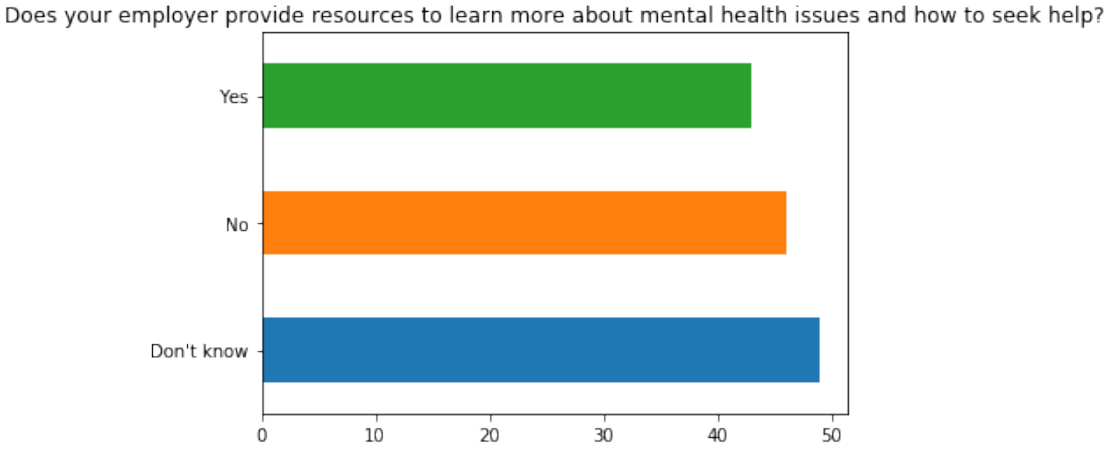
```
In [226]: # mental health being part of a wellness program
          df_ca['wellness_program'].value_counts().plot(title='Has your employer ever discusse
          df_ca['wellness_program'].value_counts()
```

```
Out[226]: No            69
          Yes           36
          Don't know    33
          Name: wellness_program, dtype: int64
```



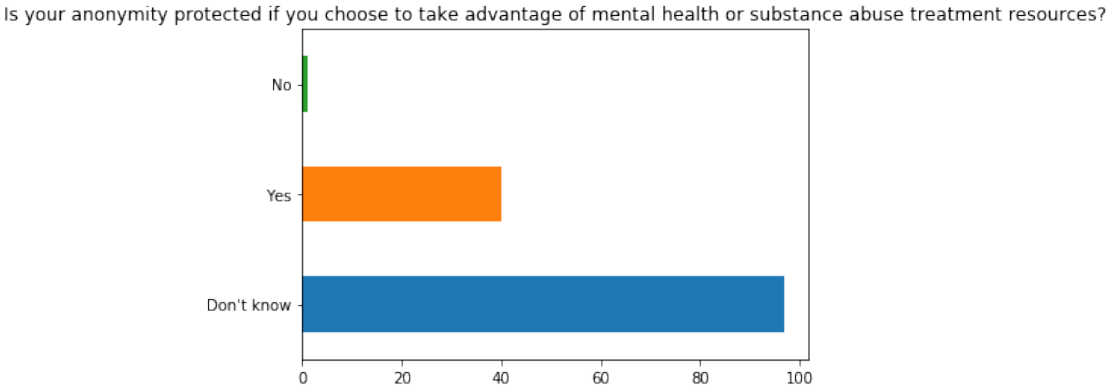
Many employees reported never having discussed mental health with their employers. Therefore, a lot of people probably are not even aware of the mental health isuues that surround them.

```
In [227]: #seeking help
          df_ca['seek_help'].value_counts().plot(title='Does your employer provide resources to
          df_ca['seek_help'].value_counts()

Out[227]: Don't know    49
          No            46
          Yes           43
          Name: seek_help, dtype: int64
```



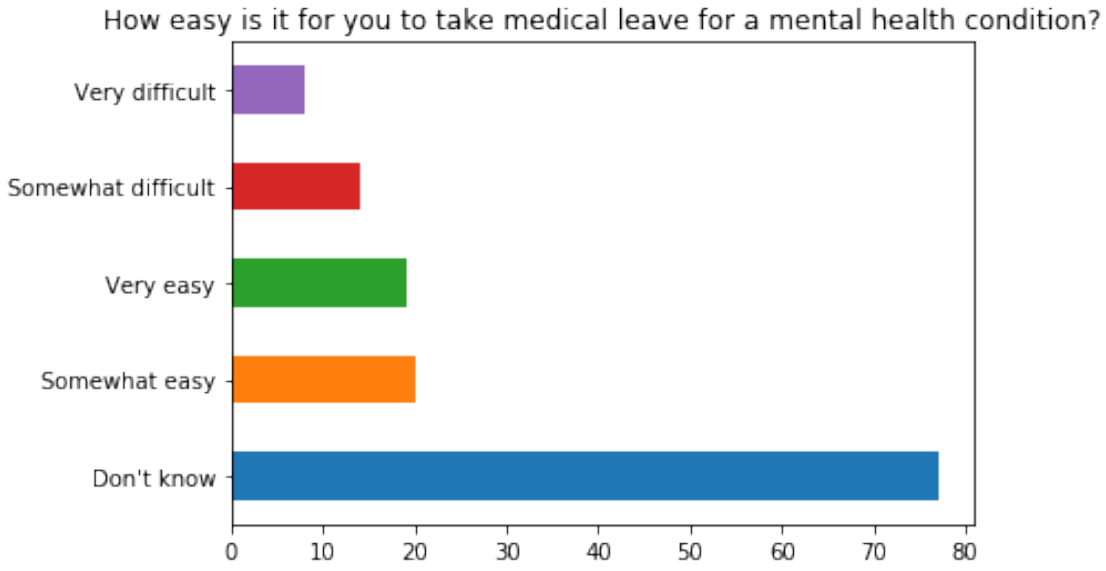
More people said "no" to their employers providing resources to learn about mental health than "yes", but even more said that they don't know. This means that the majority of employees are not able to learn more about mental health and seek help, and this is not good because when people do not have the opportunity to get the help they need or do not feel comfortable doing so, it pushes them even further into their issues.

```
In [228]: #staying anonymous about their mental health condition
          df_ca['anonymity'].value_counts().plot(title='Is your anonymity protected if you choo
          df_ca['anonymity'].value_counts()

Out[228]: Don't know    97
          Yes           40
          No             1
          Name: anonymity, dtype: int64
```

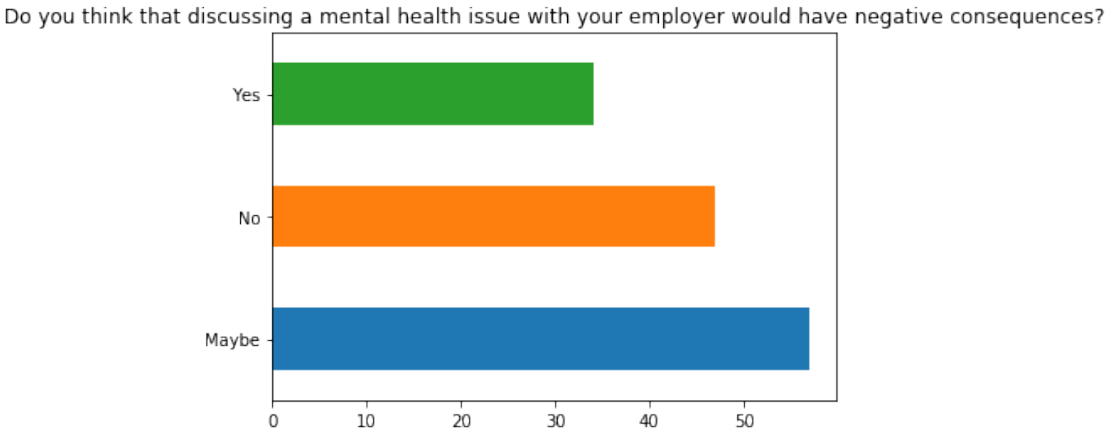


Again, so many people do not know the answers to these questions, meaning that mental health is not really being treated as something important in these workplaces.

```
In [229]: #if they can receive medical leave for their mental health
          df_ca['leave'].value_counts().plot(title='How easy is it for you to take medical leav
          df_ca['leave'].value_counts()

Out[229]: Don't know              77
          Somewhat easy           20
          Very easy               19
          Somewhat difficult      14
          Very difficult           8
          Name: leave, dtype: int64
```

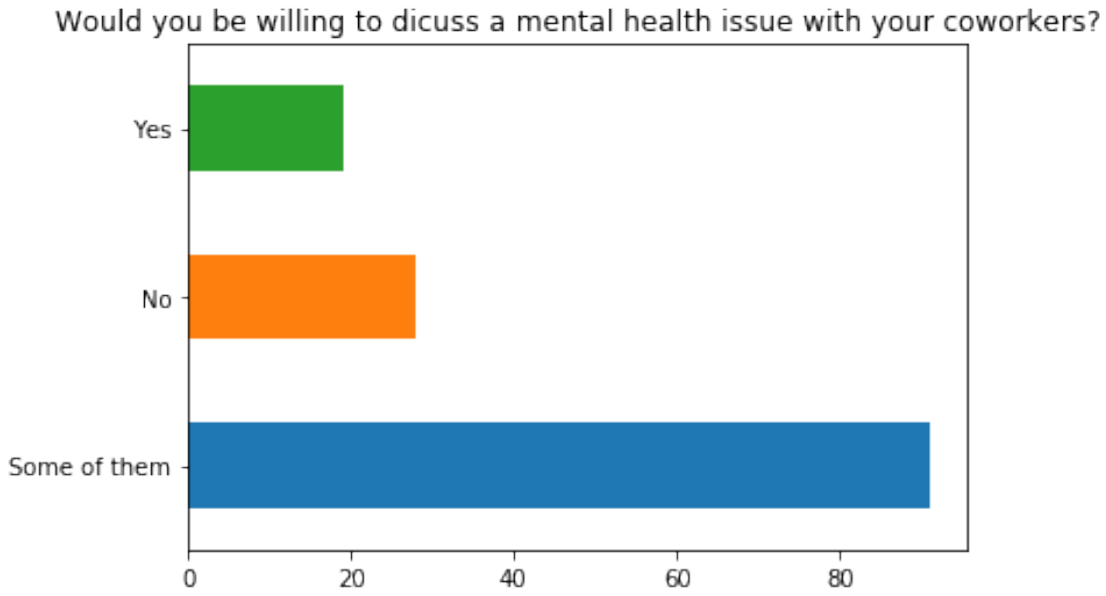


```
In [230]: #if discussing their mental health will have deprecating consequences on them
          df_ca['mental_health_consequence'].value_counts().plot(title='Do you think that discu
          df_ca['mental_health_consequence'].value_counts()

Out[230]: Maybe    57
          No       47
          Yes      34
          Name: mental_health_consequence, dtype: int64
```



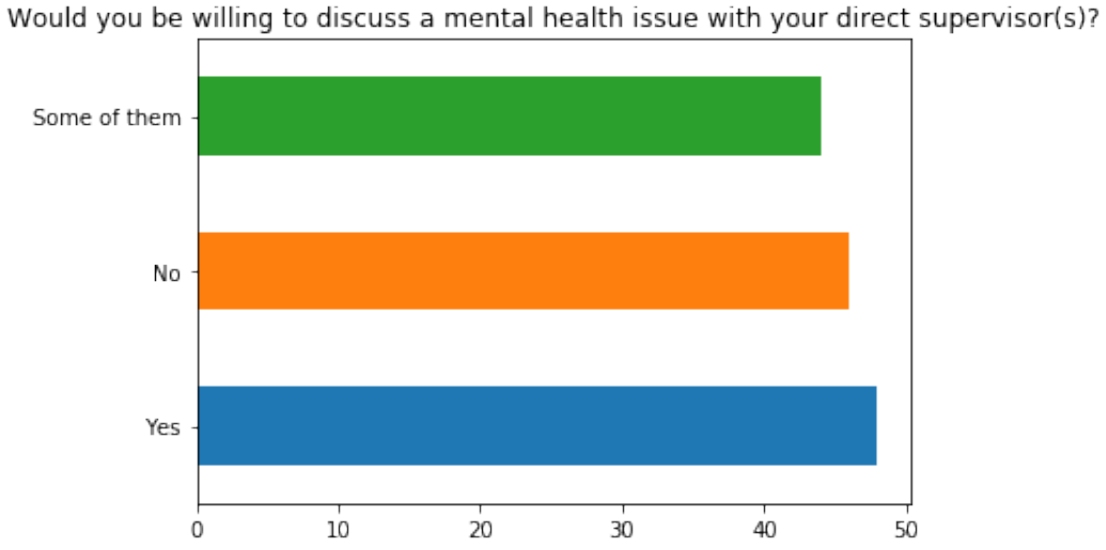
According to these results, the people who said that discussing mental health issues with their employer would or might have negative consequences are in the majority. This shows how much the work environments in tech companies do not provide a safe area where employees feel comfortable expressing these mental health issues, and in some cases, the issues could be serious, but employees feel that they cannot discuss them. This just leads to people repressing their issues, which can make them even worse.

```
In [231]: #if they are willing to discuss their mental health issues with their co-workers
          df_ca['coworkers'].value_counts().plot(title='Would you be willing to dicuss a mental
          df_ca['coworkers'].value_counts()

Out[231]: Some of them    91
          No              28
          Yes             19
          Name: coworkers, dtype: int64
```



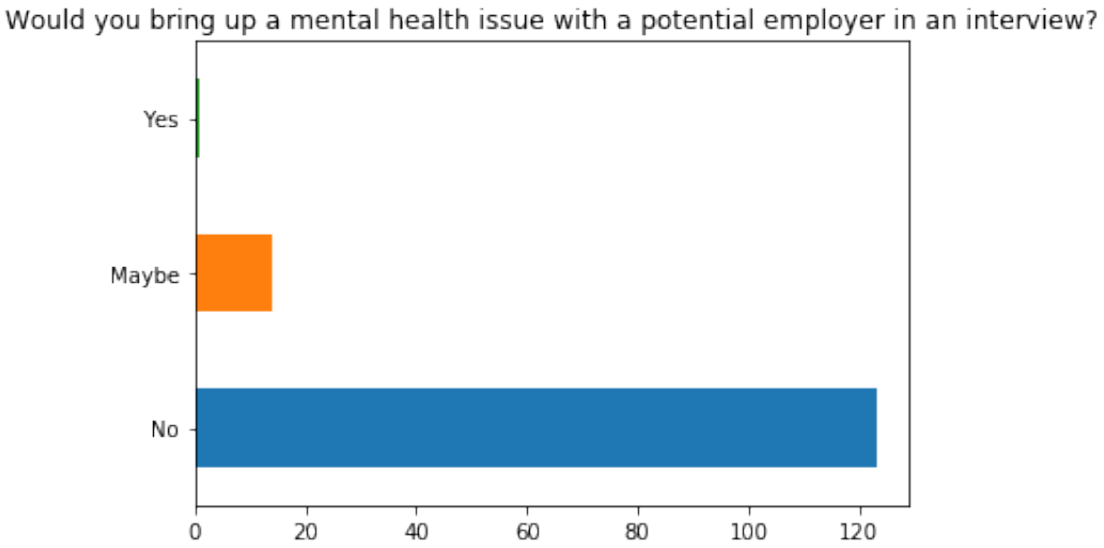
Most people would discuss their mental health issues with some of their coworkers, which is good.

```
In [232]: #are they able to talk about their mental health issues with their direct supervisor
          df_ca['supervisor'].value_counts().plot(title='Would you be willing to discuss a ment
          df_ca['supervisor'].value_counts()

Out[232]: Yes             48
          No              46
          Some of them    44
          Name: supervisor, dtype: int64
```



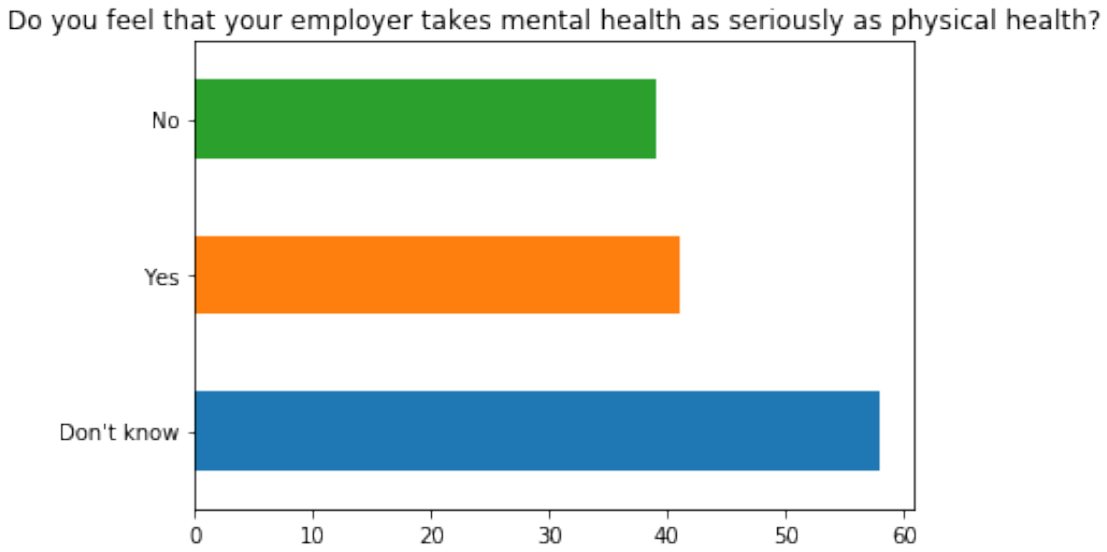
Based on these results, many employees are not willnig to discuss their mental health issues with their superiors. This is not good because it could make their supervisors unaware of their employees' health conditions and uninententionally could cause them to do things that are harmful to their employees and their mental health.

```
In [233]: #would they bring up their mental health issue with their potential employers
          df_ca['mental_health_interview'].value_counts().plot(title='Would you bring up a ment
          df_ca['mental_health_interview'].value_counts()

Out[233]: No       123
          Maybe     14
          Yes        1
          Name: mental_health_interview, dtype: int64
```

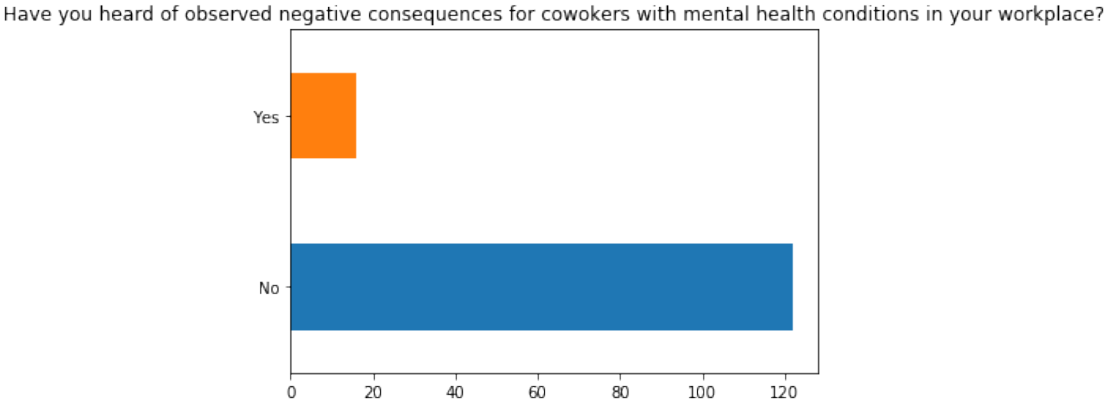


Very few people would bring up their mental health issues in job interviews. This could be because people believe that their metnal health issues would be a disadvantage and would decrease their credibility as a candidate. It makes sense for people to believe this, but it also shows that properly taking care of mental health in the workplace is just swept under the rug and does not seem to be a priority.

```
In [234]: #can mental health be taken as seriously as physical health
          df_ca['mental_vs_physical'].value_counts().plot(title='Do you feel that your employe
          df_ca['mental_vs_physical'].value_counts()

Out[234]: Don't know    58
          Yes           41
          No            39
          Name: mental_vs_physical, dtype: int64
```




```
In [235]: #observing negative consequences for coworkers who discussed their mental health con
          df_ca['obs_consequence'].value_counts().plot(title='Have you heard of observed negati
          df_ca['obs_consequence'].value_counts()

Out[235]: No     122
          Yes     16
          Name: obs_consequence, dtype: int64
```



Most employees have not heard of negative consequences for those with mental health conditions in their workplace, so it shows that either no one ever talks about it, or the people who do talk about it end up being fine and maybe even benefit from it.

As we can see from this survey from people who have mental health issues, most of them are not willing to speak about their conditions with their employers or in the workplace. Many of them are also unaware of the different kinds of benefits or resources that are open to them that can

help with mental health. This shows that employers in tech companies do not typically prioritize these issues.

In our project we wanted to see what kinds of factors might influence STEM students to develop mental health issues, and we predicted that the competitive nature of the classes they take push them to do more than they can, which carries over to the workplace as well.

11 Ethics & Privacy

The datasets we used for our project were acquired from Kaggle, an online resource hub for the data science and machine learning community, and from the UCSD CAPE website, a collection of course and professor evaluations from the University of California, San Diego. The datasets are open-source, so there was no need to seek approval to use them for our project. While there were no explicit privacy terms that needed to be complied with , we did want to keep in mind the ethics of working with data that highlight an individual's medical history, which in this case are mental health illnesses. There are indeed biases that may have influenced our data analysis, particularly in that the University of California, San Diego is a predominantly STEM-major college. This suggests a number of environmental and psychological factors that are specific to attending a school of such nature, such as having more immediate access to exceptional educational resources and that attending a STEM-centered college can either create a more competitive/supportive learning environment, depending on the attitudes that compose the student body. In order to mitigate biases that affect the grades at a STEM-centered school like UCSD, our team took a look at grade distribution data from The University of Washington, whose student body population reflects a much wider range of majors.

12 Summary of Data and Question

Using CAPE UCSD data, students in STEM classes struggle to receive good grades even if they study for at least 8 hours. Specifically, students in the mechanical engineering and chemistry majors have the highest percentage of students studying more than 8 hours a week. While, students who are in non-STEM classes still receive better grades even when they do not study more than 8 hours. This data emphasizes that STEM classes require more time and effort due to the rigors of the classes they are required to take. This may explain why STEM students are more likely to suffer from mental health issues. They are forced to spend numerous hours studying to be able to succeed in these classes.

From the Exploratory Analysis on Worst Grades data, it is certain that the highest rate of D and F letters found mostly in STEM related courses with the highest of 0.186066 in Intermediate Algebra. Such as, Mathematics and Biology related courses have more than half of the top 100 courses with highest rate with D and F letters. While, the highest rate of A letter found mostly in Humanity and Performing Arts. It is understandable that STEM courses are often difficult and challenging compare to other courses. Therefore, the rate for worst grades commonly results in STEM related courses.

From the FiveThirtyEight College Major Graduations data, STEM related majors do not have a relatively high unemployment rate in comparison to the individuals who did not receive a graduate degree in the same field. This dataset reveal that people in the STEM field does not necessarily have to obtain a graduate degree to get a job, because date emphasizes that their employment rates and their unemployment rates does not show a significant difference. It was not significant

enough to state that obtaining a graduate degree necessarily meant that an individual's likelihood of being employed is greater than someone who does not have their graduate degree.

From the Mental Health and Tech data, the categorized questions and answers presents the prevalence of mental health issues as well as its' benefits in the tech industry. The dataset was gathered from all around the world, and our group cleaned up the dataset to only show the data that shows individuals that work in tech companies based in California. Based on this data, it revealed that around 10 percent of employers were experiencing mental health problems while working in these companies. Employees in these companies mention that their company does provide mental health benefits, which allows them to get the help they need if they decide to seek help.

13 Results & Conclusions

To recap, the research question guiding our team's data analysis is: What academic and social factors do STEM (Science, Technology, Engineering, and Mathematics) students experience during undergrad that contribute to mental-health illnesses in their professional careers? Our team our hypothesized that STEM students typically experience a higher number of study hours per week, are more prone to backlogging, and show extreme distress about their grades in comparison to their peers. There seems to be a discrepancy in the relationship between the number of hours spent studying and the average grade received–the data indicates that students in STEM classes will not achieve average grades better than a B- (3.3) despite spending at least 8 hours a week studying for that class, whereas students in non-STEM will achieve a B- about half the time when they study for the same 8 hours. Ultimately, this analysis refutes the concept that "the more one puts in, the more one will get out," which translates into the motivation that " if a student studies more, then he/she will get better grades." It is human nature to find it disappointing and discouraging when one's efforts to excel do not necessarily culminate into tangible evidence of success. In addition, this expectation of success when putting in hard work extends further into the careers of students who graduated with STEM degrees. Another part of our analysis highlights that STEM majors do not necessarily have lower unemployment rates than non-STEM majors, which again, can be frustrating for those who choose a more intensive major to increase their job prospects post-grad. Moreover, our analysis reveals that while employees in the tech industry recognize the effects of their mental health in the workplace, it is clear that the mental health is not typically acknowledged or addressed by employers. The fast-paced and rigorous nature of the tech industry bolsters the notion that has been ingrained into STEM students since college that "the more one puts in, the more one will get out," perpetuating a culture of high-expectation and high-stress which are two of many factors that contribute to mental wellness.

14 Limitations

It is possible that some of our data does not exactly represent everything accurately. For example, for the CAPE data, we got the data from a certain selection of classes. That one part may not accurately represent all STEM classes as a whole, and some majors may be over or under represented. When we dropped the rows with NaN in the CAPE data, we lost some data, and that may also have affected the results. For the survey on mental health in the workplace, there is also a possibility that the results collected are only from certain kinds of places in California where certain kinds of answers are more common, which may have skewed the data one way or another. We did what we could based on the data that we had and analyzed accordingly.

15 Impact on Society

Mental health in STEM fields is something that society should be more aware of. Based on what we saw from our data analysis, STEM students push themselves to study a lot and still don't always get good grades, which can push them into a place of stress and mental instability. They then bring these issues into the workplace where there is a similar kind of pressure and continue to suffer from these mental health conditions. Mental health should be a more common concern among the public because it is a real thing that affects people negatively, and we should be actively trying to find ways to help people with these conditions, rather than being oblivious and making it even worse.