Project Coursera

Increase enrollment in courses

Group 5

Diego Carlos, Paul Chen, Michael Jonelis, Jay Lee

Our Team

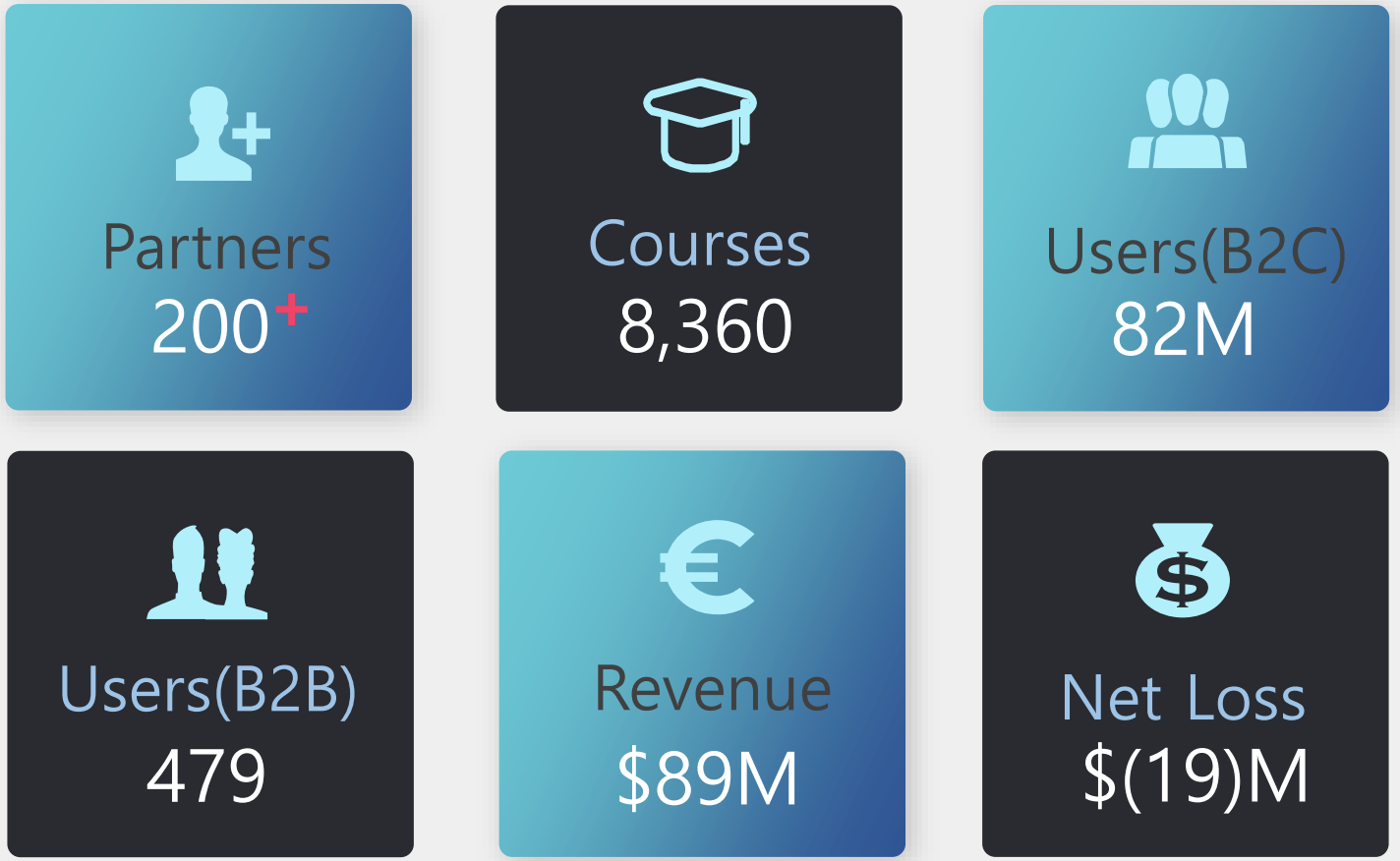
Paul Chen

Diego Carlos

Michael Jonelis

Jay Lee

Company Introduction



Overview of the case

Question: How can we increase enrollment?

Dataset

- **891 records of course information such as course title, course organization, certification type, rating, difficulty and number of enrollment**

Approaches

- **We explored its dataset to find elements that attract more enrollment**
- **It will be helpful for Coursera to determine which courses to launch within limited resources**

Dataset

```
1 df.head(5)
```

	Unnamed: 0	course_title	course_organization	course_Certificate_type	course_rating	course_difficulty	course_students_enrolled
0	134	(ISC)² Systems Security Certified Practitioner...	(ISC)²	SPECIALIZATION	4.7	Beginner	5.3k
1	743	A Crash Course in Causality: Inferring Causal...	University of Pennsylvania	COURSE	4.7	Intermediate	17k
2	874	A Crash Course in Data Science	Johns Hopkins University	COURSE	4.5	Mixed	130k
3	413	A Law Student's Toolkit	Yale University	COURSE	4.7	Mixed	91k
4	635	A Life of Happiness and Fulfillment	Indian School of Business	COURSE	4.8	Mixed	320k

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Unnamed: 0               891 non-null    int64
 1   course_title             891 non-null    object
 2   course_organization      891 non-null    object
 3   course_Certificate_type  891 non-null    object
 4   course_rating            891 non-null    float64
 5   course_difficulty        891 non-null    object
 6   course_students_enrolled 891 non-null    object
dtypes: float64(1), int64(1), object(5)
memory usage: 48.9+ KB
```

Issues with Data

```
1 df.head(5)
```

	Unnamed: 0	course_title	course_organization	course_Certificate_type	course_rating	course_difficulty	course_students_enrolled
0	134	(ISC)² Systems Security Certified Practitioner...	(ISC)²	SPECIALIZATION	4.7	Beginner	5.3k
1	743	A Crash Course in Causality: Inferring Causal...	University of Pennsylvania	COURSE	4.7	Intermediate	17k
2	874	A Crash Course in Data Science	Johns Hopkins University	COURSE	4.5	Mixed	130k
3	413	A Law Student's Toolkit	Yale University	COURSE	4.7	Mixed	91k
4	635	A Life of Happiness and Fulfillment	Indian School of Business	COURSE	4.8	Mixed	320k

1 3 4 2 3

Category

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Unnamed: 0               891 non-null    int64
 1   course_title             891 non-null    object
 2   course_organization      891 non-null    object
 3   course_Certificate_type  891 non-null    object
 4   course_rating            891 non-null    float64
 5   course_difficulty        891 non-null    object
 6   course_students_enrolled 891 non-null    object
dtypes: float64(1), int64(1), object(5)
memory usage: 48.9+ KB
```

1

1. Change the data types and drop Unnamed column
2. Eliminate 'k', 'm' and change them into actual numbers
3. Add category information for each course
4. Identify organization type(universities/companies)

Data Cleaning

Step 01

Change data types
&
Drop a column

Step 02
Change 'k' and 'm' into actual numbers.

Step 03
Added categories

Step 04
Identify organization types

```
#correct the types
df_temp[['course_Certificate_type', 'course_difficulty']] = \
df_temp[['course_Certificate_type', 'course_difficulty']].astype('category')

#drop useless columnmn
data1=data1.drop(labels='Unnamed: 0',axis=1)
```

Data Cleaning

Step 01
Change data types &
Drop a column

Step 02

Change 'k' and 'm' into actual numbers

Step 03
Added categories

Step 04
Identify organization types

```
#calculate enrollment numbers
df_temp['new_course'] = df_temp['course_students_enrolled'].str.replace('k', '')
df_temp['new_course'] = df_temp['new_course'].str.replace('m', '')
df_temp['new_course'] = df_temp['new_course'].astype('float')
df_temp['new2_course'] = np.where(df_temp['course_students_enrolled'].str.find('k') != -1,1000,1000000)
```

```
#after calculation, drop temp tables
df_temp['num_enrollment']=df_temp['new_course'] * df_temp['new2_course']
df_temp['num_enrollment']=df_temp['num_enrollment'].astype('int')
df_temp.drop(labels = ['course_students_enrolled', 'new_course', 'new2_course'], axis = 1, inplace = True)
```

Data Cleaning

Step 01
Change data types &
Drop a column

Step 02
Change 'k' and 'm' into
actual numbers

Step 03

Added categories

Step 04
Identify organization
types

Manually Added

Data Cleaning

Step 01
Change data types & Drop a column

Step 02
Change 'k' and 'm' into actual numbers

Step 03
Added categories

Step 04

Identify organization types

```
conditions=[
    (df_temp['course_organization'].str.find('Institute for the Future')!=-1),
    (df_temp['course_organization'].str.find('École')!=-1),
    (df_temp['course_organization'].str.find('Universidad')!=-1),
    (df_temp['course_organization'].str.find('College')!=-1),
```

```
values=['Company','University','University','University','University','University'

df_temp['type_org']=np.select(conditions,values)
```

After Cleaning the dataset

```
1 df.head(5)
```

	Unnamed: 0	course_title	course_organization	course_Certificate_type	course_rating	course_difficulty	course_students_enrolled
0	134	(ISC)² Systems Security Certified Practitioner...	(ISC)²	SPECIALIZATION	4.7	Beginner	5.3k
1	743	A Crash Course in Causality: Inferring Causal...	University of Pennsylvania	COURSE	4.7	Intermediate	17k
2	874	A Crash Course in Data Science	Johns Hopkins University	COURSE	4.5	Mixed	130k
3	413	A Law Student's Toolkit	Yale University	COURSE	4.7	Mixed	91k
4	635	A Life of Happiness and Fulfillment	Indian School of Business	COURSE	4.8	Mixed	320k

```
1 df_temp.head(5)
```

	course_title	course_organization	course_Certificate_type	course_rating	course_difficulty	Category	num_enrollment	type_org
0	(ISC)² Systems Security Certified Practitioner...	(ISC)²	SPECIALIZATION	4.7	Beginner	Information Technology	5300	Company
1	A Crash Course in Causality: Inferring Causal...	University of Pennsylvania	COURSE	4.7	Intermediate	Business	17000	University
2	A Crash Course in Data Science	Johns Hopkins University	COURSE	4.5	Mixed	Data Science	130000	University
3	A Law Student's Toolkit	Yale University	COURSE	4.7	Mixed	Social Sciences	91000	University
4	A Life of Happiness and Fulfillment	Indian School of Business	COURSE	4.8	Mixed	Health	320000	University

Business Questions

1. Does course rating impact course enrollment?

2. Does difficulty impact rating or enrollment?

3. Do universities or companies have higher ratings and enrollment?

4. Which type of certificates has the highest ratings and enrollment?

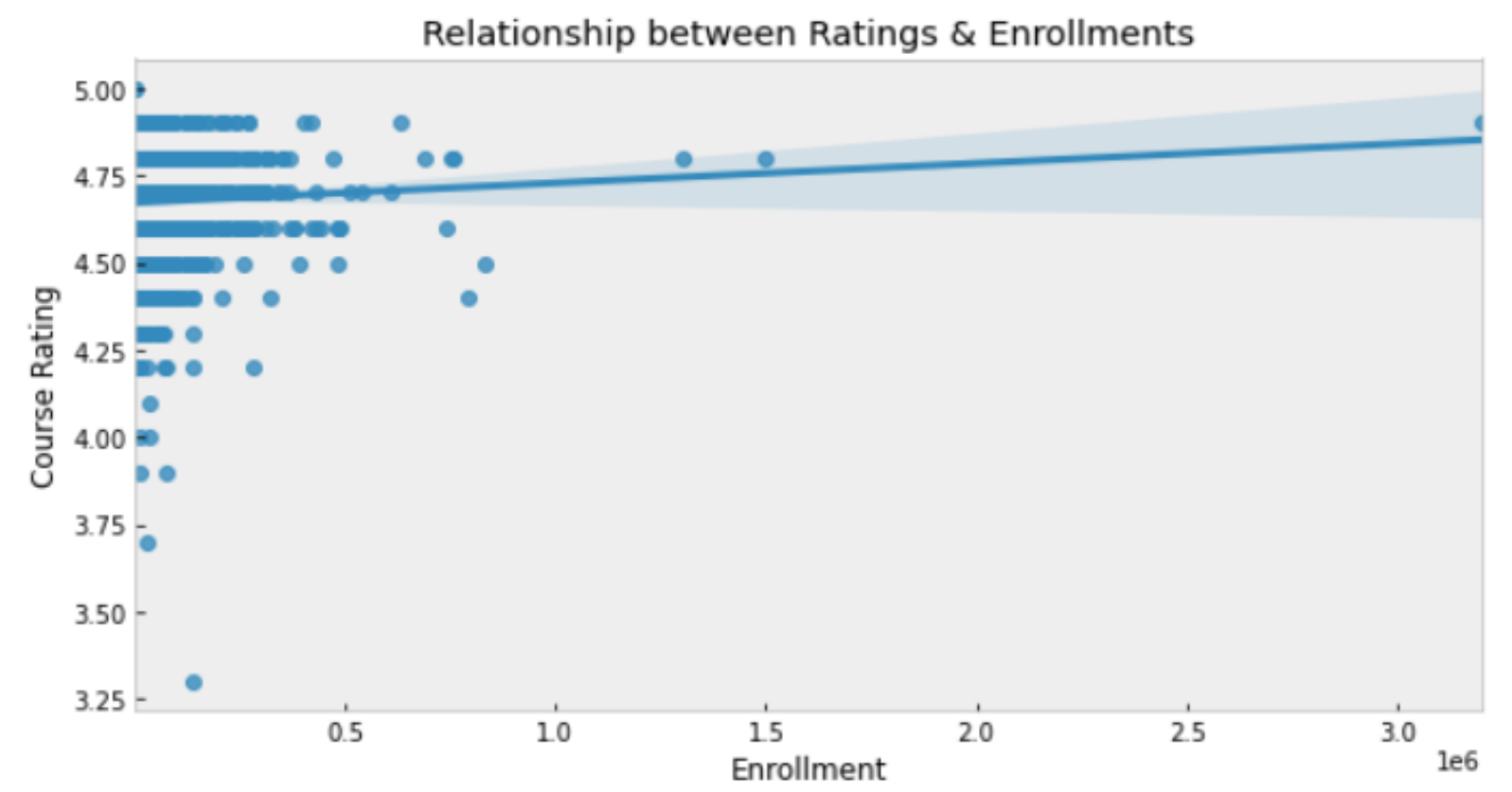
5. Which subject categories have the highest average enrollment?

6. Which companies and universities have the highest average enrollment?

7. Which organizations preform best in the most popular categories?

1. Does course rating impact course enrollment?

Positive relationship between ratings and enrollment



P-Value
0.013

Ratings
Enrollment

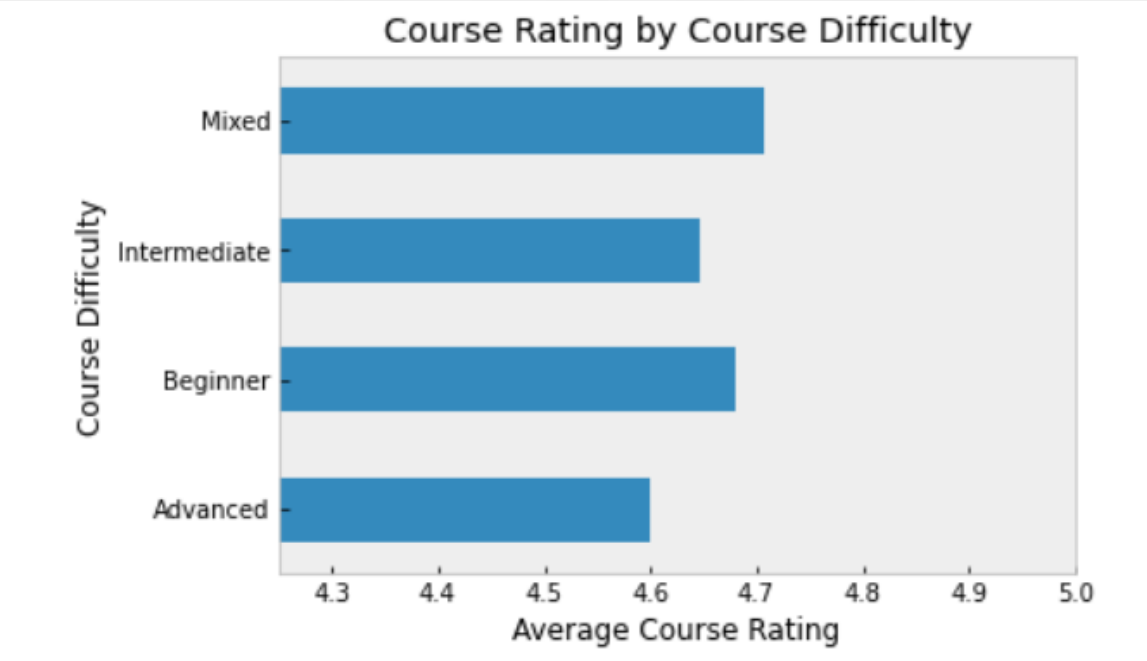
```
plt.figure(figsize=(15,10))
sns.regplot(x="num_enrollment", y="course_rating", data=df_temp)
plt.xlabel('Numbers of enrollment')
plt.ylabel('Course Rating')
plt.title('Relationship between Ratings & Enrollments')
plt.grid()
```

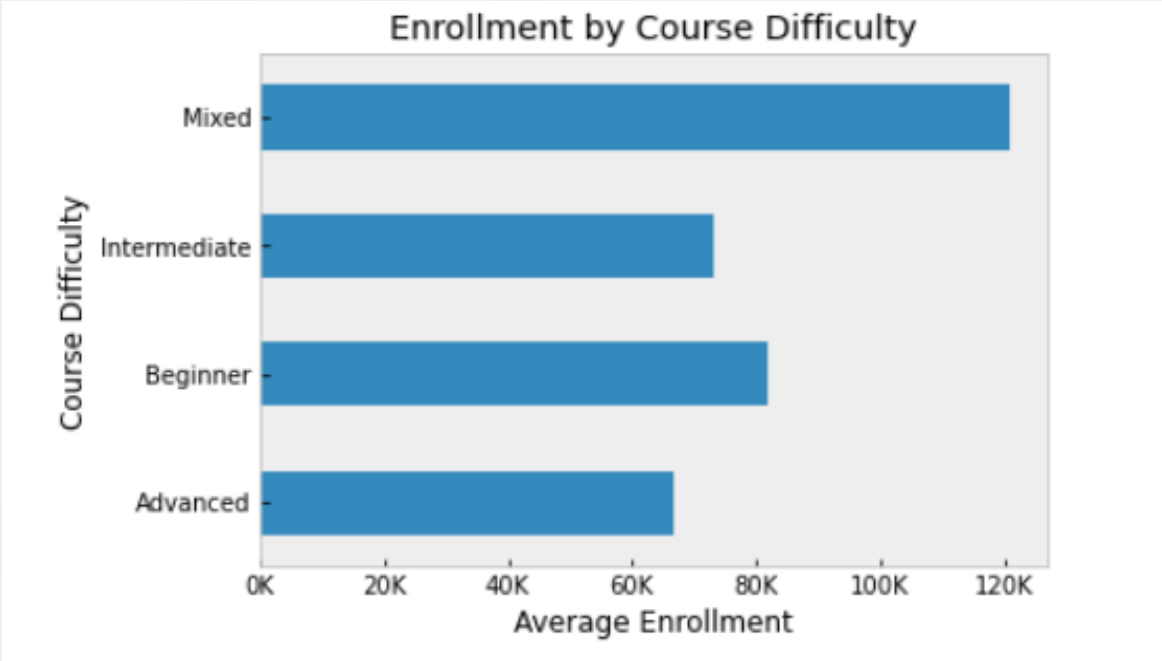
$$Enrollment_i = \beta_0 + \beta_1 Rating_i + \sum \gamma_i CertificationType_i + \sum \delta_i Difficulty_i + \sum \theta_i Category_i + \sum \rho_i Institution_i + \varepsilon_i$$

2. Does difficulty impact ratings or enrollment?

Difficulty has a relationship with both ratings and enrollment

One-way **ANOVA** shows that enrollment and ratings vary between difficulty levels

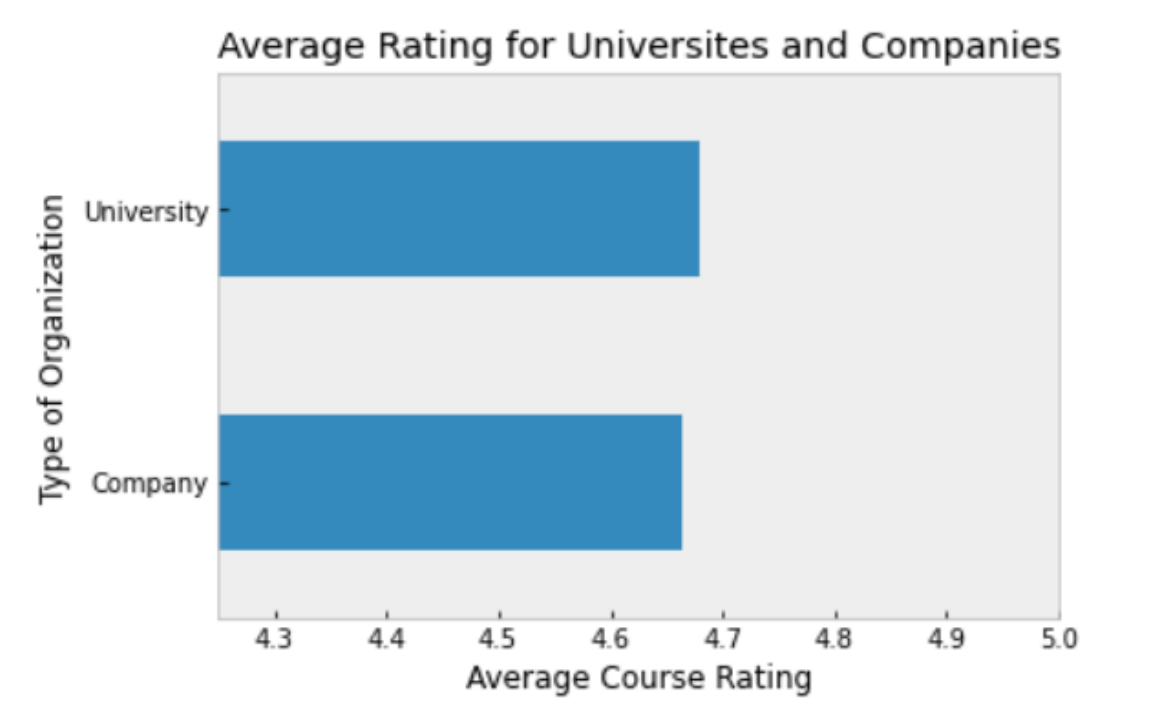


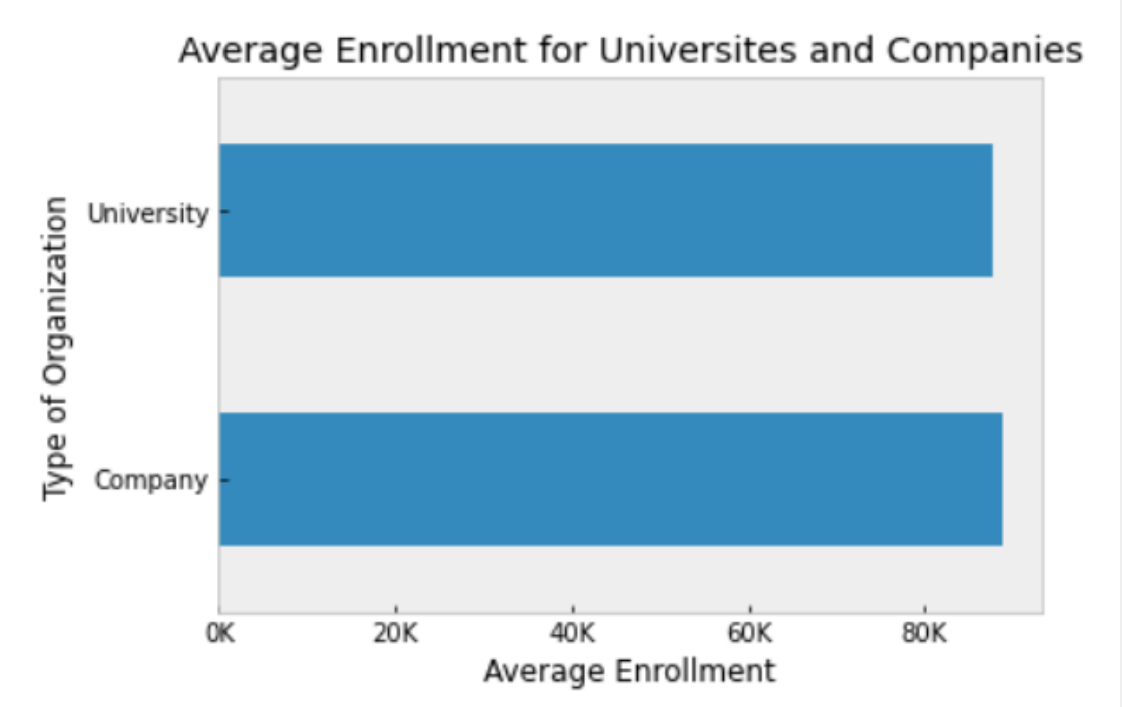


3. Do universities or companies have higher ratings and enrollment?

It does not matter!

One-way **ANOVA** does not show that enrollment nor ratings vary between universities and companies

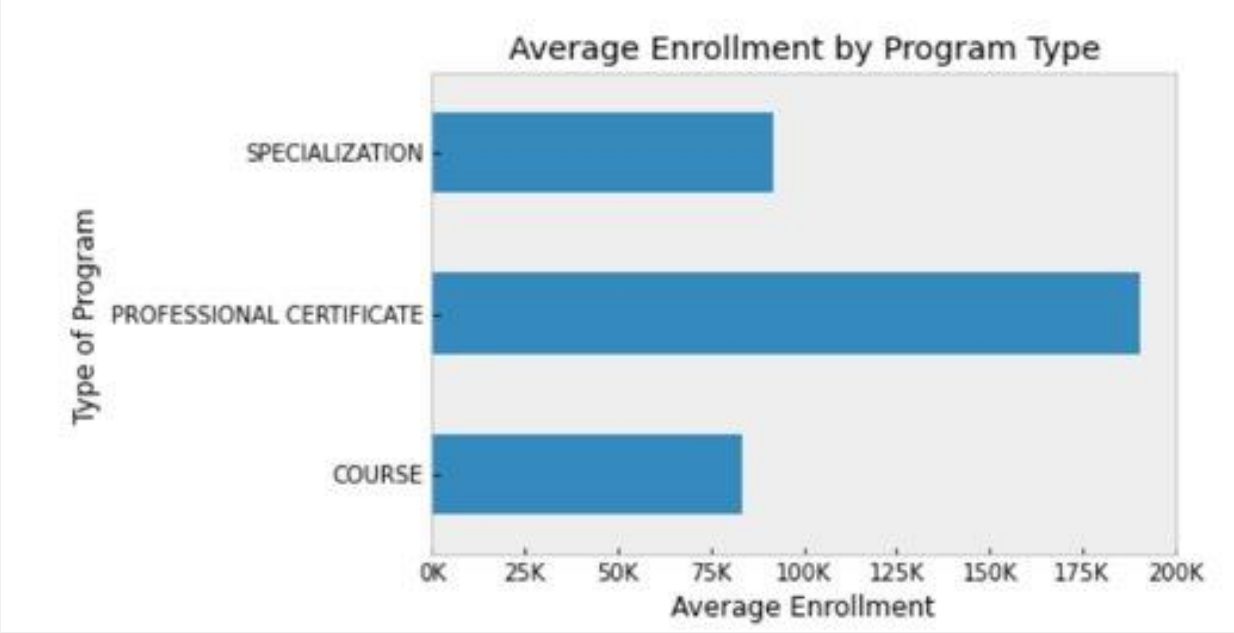





```
plt.figure(figsize=(8,4))
aggreg_data4.plot(kind='barh', x='type_org',y='num_enrollment',legend=None)\
.get_xaxis().set_major_formatter(tkr.FuncFormatter(lambda x, pos: '{:,.0f}'.format(x/1000) + 'K'))
#plt.xlim([4.25, 5.00])
plt.xlabel('Average Enrollment')
plt.ylabel('Type of Organization')
plt.title('Average Enrollment for Universites and Companies')
plt.grid()
```

4. Which types of courses have the highest ratings and enrollment?

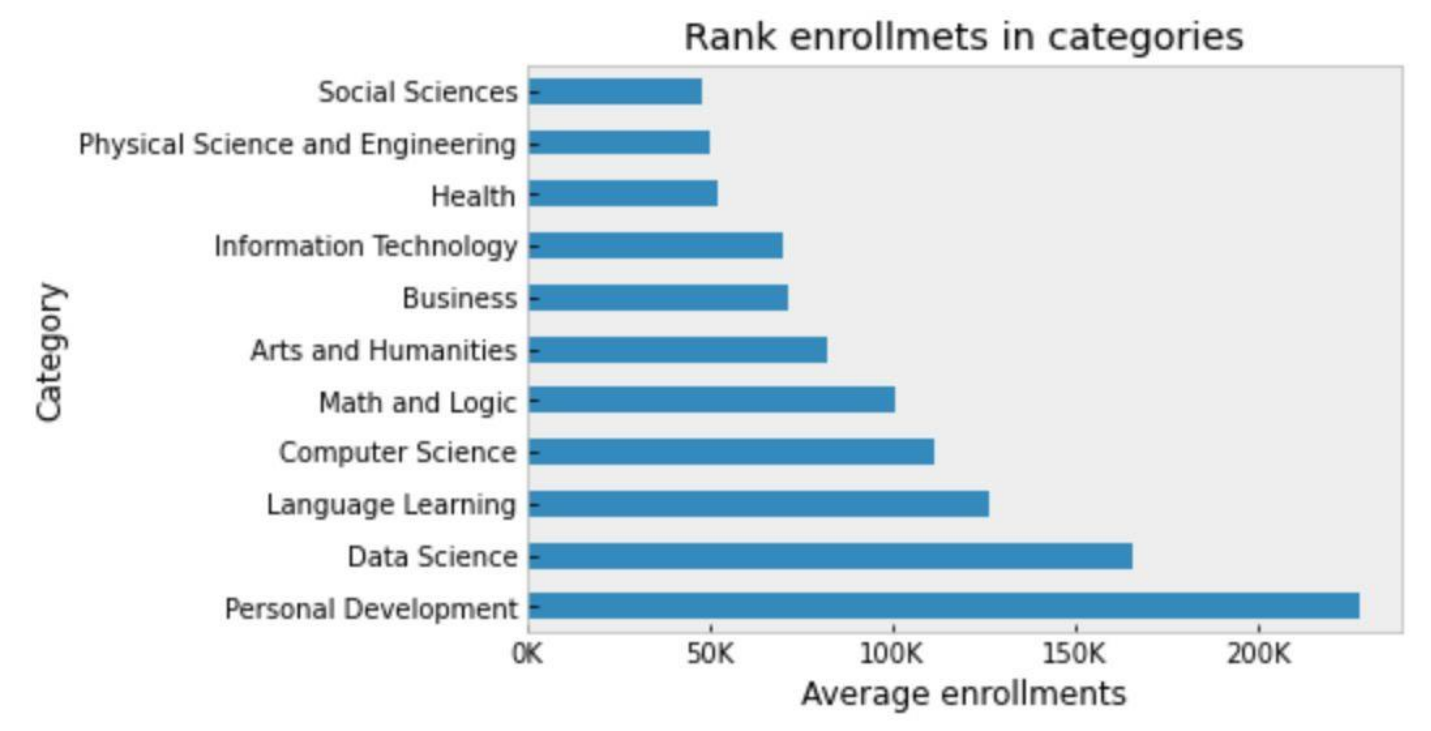
Course type does not impact enrollment, but does impact ratings




```
aggreg_data=df_temp.groupby('course_Certificate_type')['course_rating'].mean()
aggreg_data=pd.DataFrame(aggreg_data)
aggreg_data=aggreg_data.reset_index()

plt.figure(figsize=(8,4))
aggreg_data.plot(kind='barh', x='course_Certificate_type',y='course_rating',legend=None)
plt.xlabel('Average Rating')
plt.xlim([4.00,5.00])
plt.ylabel('Type of Program')
plt.title('Average Rating by Program Type')
plt.grid()
```

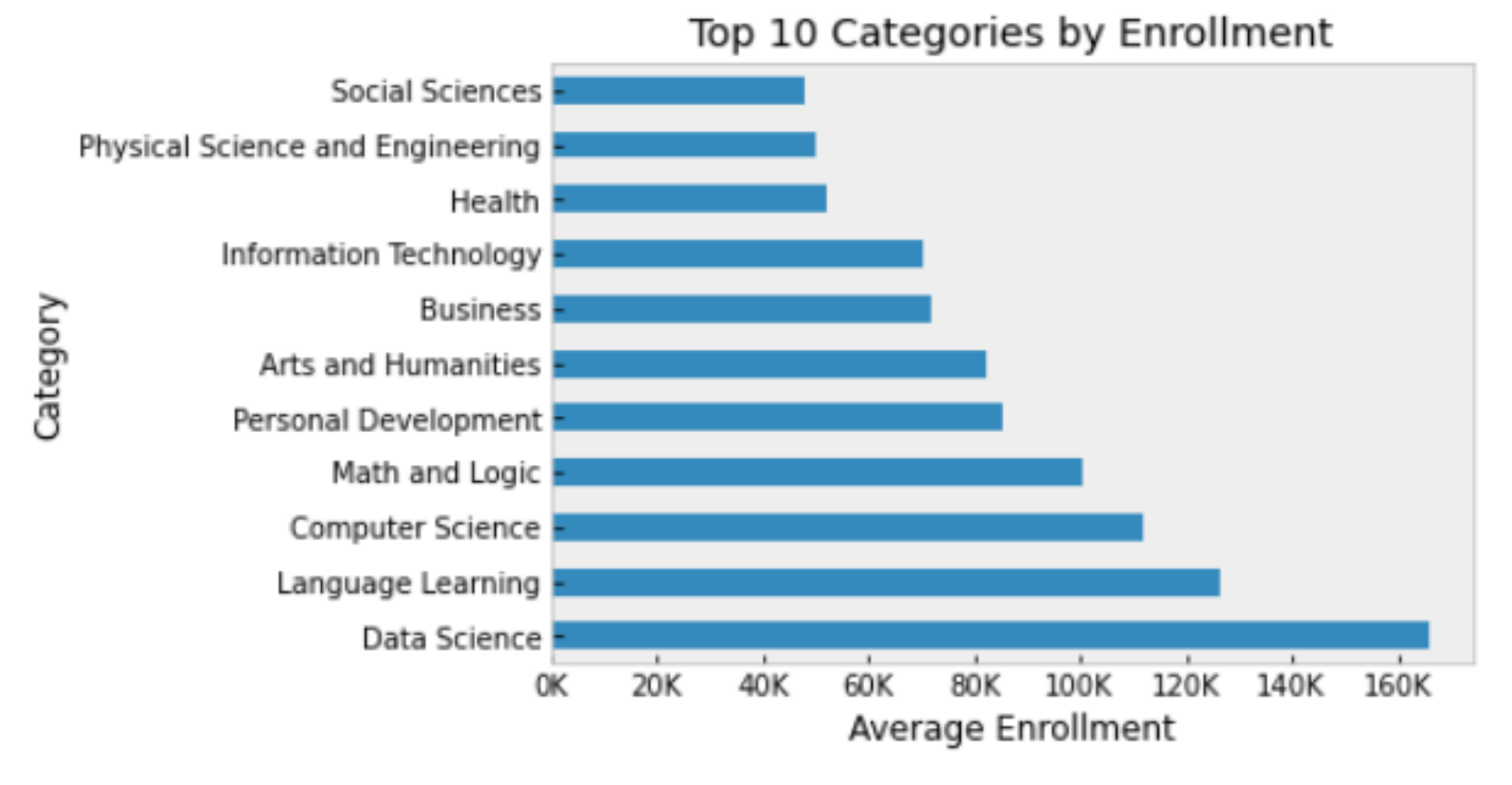
5. Which subject categories have the highest average enrollment?

Personal Development has a very high average enrollment



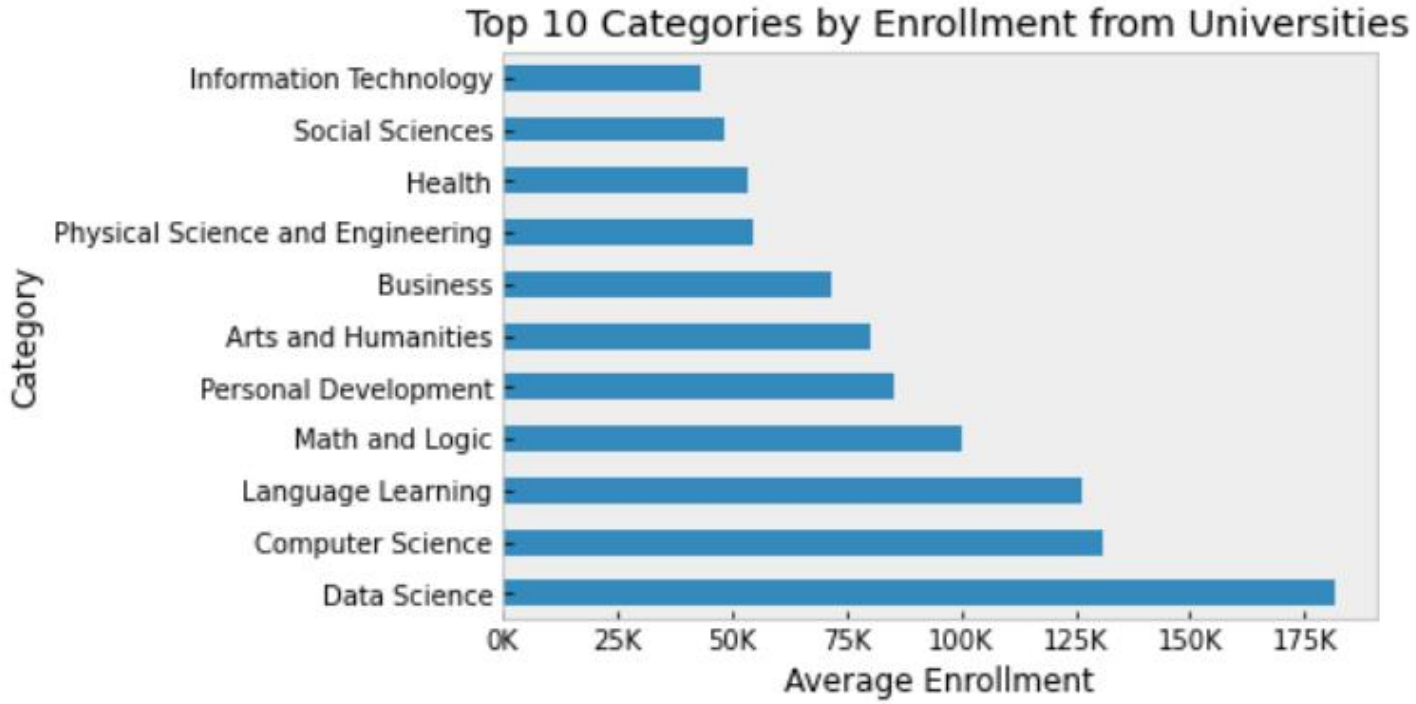

```
#plot category by enrollments
df_category_enroll.plot(kind='barh', x='Category',y='num_enrollment',legend=None).get_xaxis()\
.set_major_formatter(tkr.FuncFormatter(lambda x, pos: '{:,.0f}'.format(x/1000) + 'K'))
plt.grid()
plt.ylabel("Category")
plt.xlabel("Average enrollments")
plt.title("Rank enrollmets in categories")
```

After removing *The Science of Wellbeing*, Data Science is the most popular category



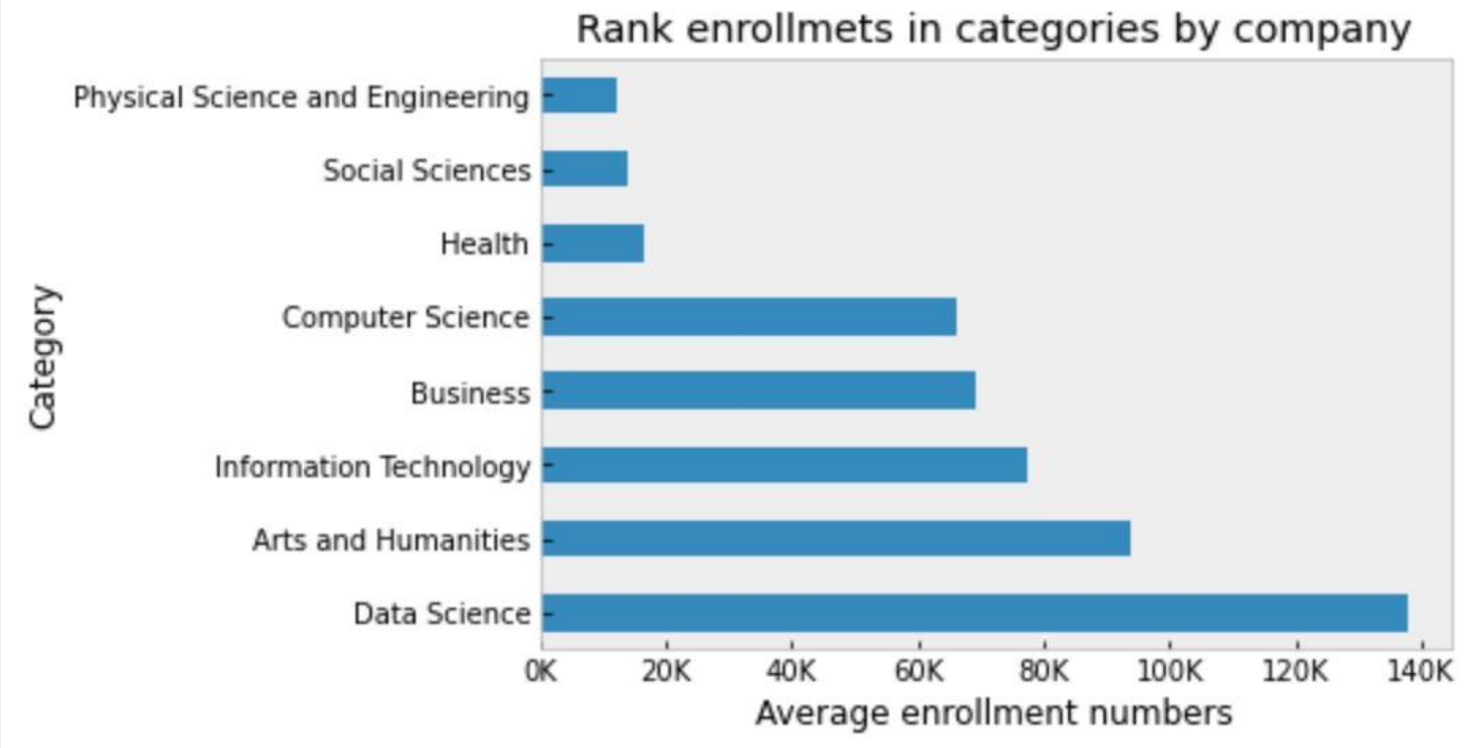

```
emp.value_counts(['Category'])

gory
ness                                294
 Science                            120
th                                  118
uter Science                        110
al Sciences                          56
 and Humanities                      49
rmation Technology                   46
uage Learning                        41
ical Science and Engineering         35
onal Development                     17
 and Logic                            5
```

Data Science is the most popular in Universities



Category	
Business	278
Health	114
Computer Science	77
Data Science	77
Social Sciences	55
Arts and Humanities	42
Language Learning	41
Physical Science and Engineering	31
Personal Development	16
Information Technology	10
Math and Logic	5

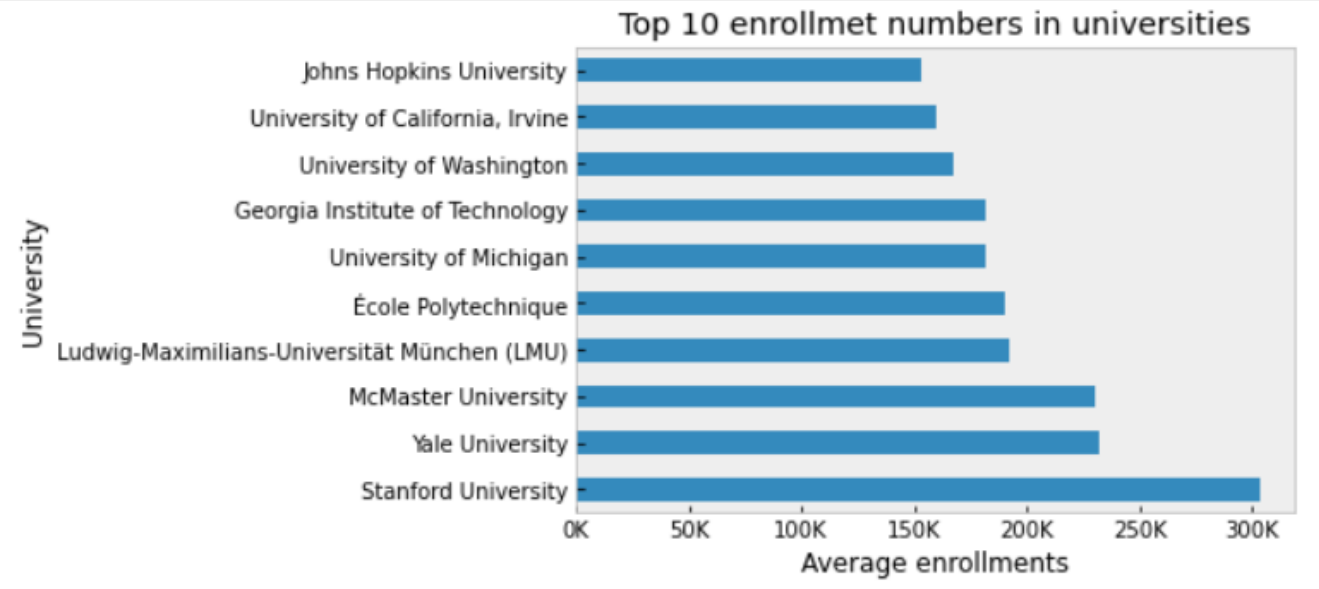
Data Science is the most popular in Companies

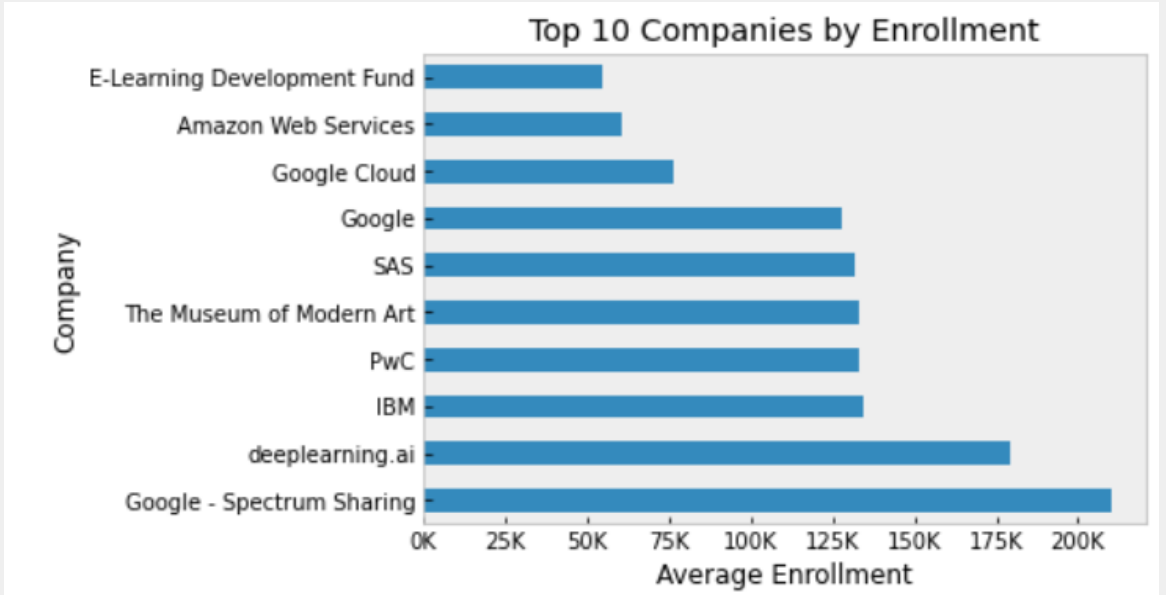


Category	
Data Science	43
Information Technology	36
Computer Science	33
Business	16
Arts and Humanities	7
Health	4
Physical Science and Engineering	4
Social Sciences	1
Language Learning	0
Math and Logic	0
Personal Development	0

6. Which companies and universities have the highest average enrollment ?

Stanford and Google






```
#plot the top 10 enrollments by university
df_uni10.plot(kind='barh', x='course_organization',y='num_enrollment',legend=None)\
.get_xaxis().set_major_formatter(tkr.FuncFormatter(lambda x, pos: '{:,.0f}'.format(x/1000) + 'K'))
plt.grid()
plt.ylabel("University")
plt.xlabel("Average enrollments")
plt.title("Top 10 enrollment numbers in universities")
plt.show()
plt.savefig('Q6 Top 10 enrollmets in universities.jpg')
#for question 6
```

7. Which organizations perform best in the most popular categories?

Exceptional Providers

Data Science

Computer Science

Language Learning

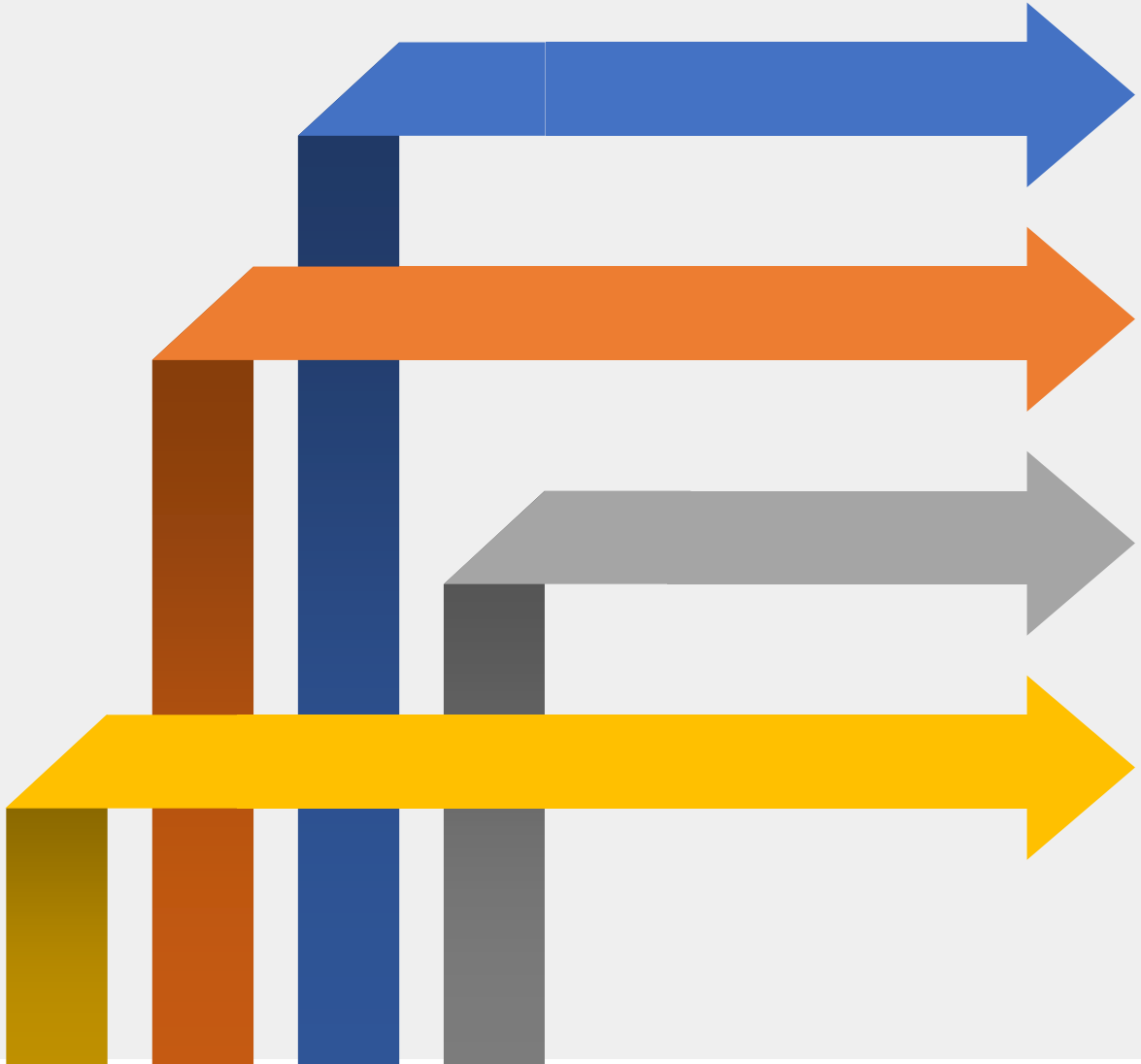
IT

Personal Development

Conclusions

- Work with less popular categories
- Targeted Ads for most popular stuff
- Work with providers for new courses

Improvements



Time Series data

Tracking subscribers over time would provide more valuable insights

Sales/ Costs data

Construct models to optimize what kind of courses to launch, which providers to partner with, and how to market them

Recent Data and more random samples

The most recent data can provide insights for a post-covid world. Random sampling is import for casual inference

Individual User Data

Analyzing which courses are taken by the same users could provide insight into which subscribers are most profitable

What we have tried

Data scrapping

BeautifulSoup

```
from bs4 import BeautifulSoup
import requests
#get info from website
url='https://www.coursera.org/courses'
page = requests.get(url)
soup = BeautifulSoup(page.text, 'lxml')

#find course title
y = soup.find_all('h2')
#<h2 class="cds-111 card-title css-1fkiswk cds-113">Google Data Analytics</h2>
print(y)
```

```
[]
[<h2 class="cds-111 rc-NumberOfResultsSection css-123aj4z cds-113" data-e2e="NumberOfResultsSection"><span>Showing 8340 total r
esults</span></h2>, <h2 class="cds-111 css-7rz9ct cds-113">What Coursera Has to Offer</h2>, <h2 class="sr-only">Coursera Footer
</h2>]
```

selenium

```
from selenium import webdriver
diverPath='D:\chromedriver_win32\chromedriver.exe'
browser=webdriver.Chrome(diverPath)

url='https://www.coursera.org/courses'
browser.get(url)
x = browser.find_elements_by_tag_name('h2')
for data in range(len(x)):
    print(x[data].text)
```

```
No results found for your search
What Coursera Has to Offer
Coursera Footer
```

Thanks for Watching

Q&A

Appendix

ANOVA-Course Type

Rating

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Course	582	2739.5	4.7070	0.0230
Specialization	297	1371.6	4.6182	0.0282
Professional Certificate	12	56.4	4.7000	0.0145

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.5591	2	0.7796	**31.6629**	**5.22E-14**	**3.0059**
Within Groups	21.8629	888	0.0246			
Total	23.4220	890				

Enrollment

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Course	582	51,131,300	87,854	38,192,986,581
Specialization	297	27,262,200	91,792	23,241,667,434
Professional Certificate	12	2,288,400	190,700	24,097,392,727

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	125,047,099,668	2	62,523,549,834	**1.8927**	**0.1513**	**3.0059**
Within Groups	29,334,730,083,991	888	33,034,605,950			
Total	29,459,777,183,659	890				

ANOVA-Institution Type

Rating

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Company	67	313	5	0
University	747	3,496	5	0

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0	1	0	**0.3705**	**0.5429**	**3.8529**
Within Groups	22	812	0			
Total	22	813				

Enrollment

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Company	67	5,970,600	89,113	14,094,443,605
University	747	67,882,500	90,873	36,710,063,881

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	190,469,968	1	190,469,968	**0.0055**	**0.9411**	**3.8529**
Within Groups	28,315,900,000,000	812	34,871,848,440			
Total	28,316,100,000,000	813				

ANOVA-Course Difficulty

Rating

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Advanced	19	87	4.60	0.0378
Beginner	487	2,280	4.68	0.0201
Intermediate	198	920	4.65	0.0357
Mixed	187	881	4.71	0.0292

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.4908	3	0.1636	**6.3283**	**0.0003**	**2.6149**
Within Groups	22.9312	887	0.0259			
Total	23.4220	890				

Enrollment

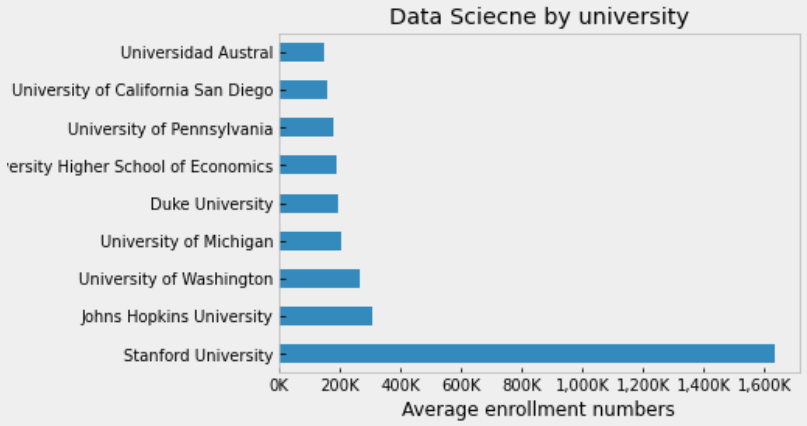
SUMMARY

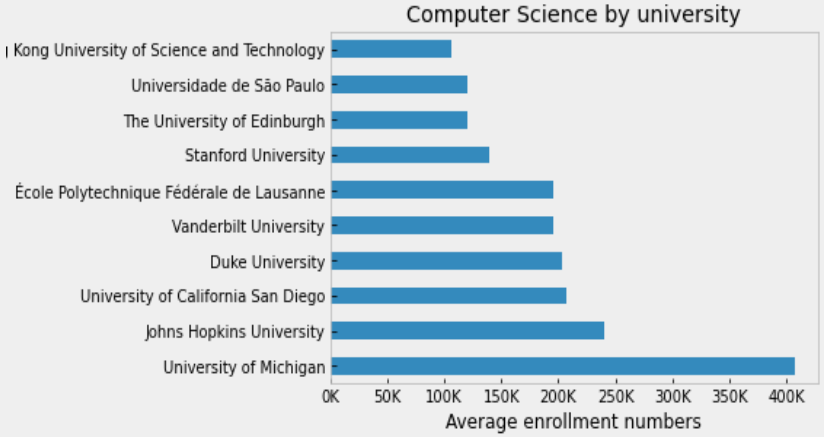
Groups	*Count*	*Sum*	*Average*	*Variance*
Advanced	19	1,264,400	66,547	6,767,071,520
Beginner	487	39,921,800	81,975	16,197,849,083
Intermediate	198	14,506,300	73,264	10,858,919,469
Mixed	187	24,989,400	133,633	101,471,000,000

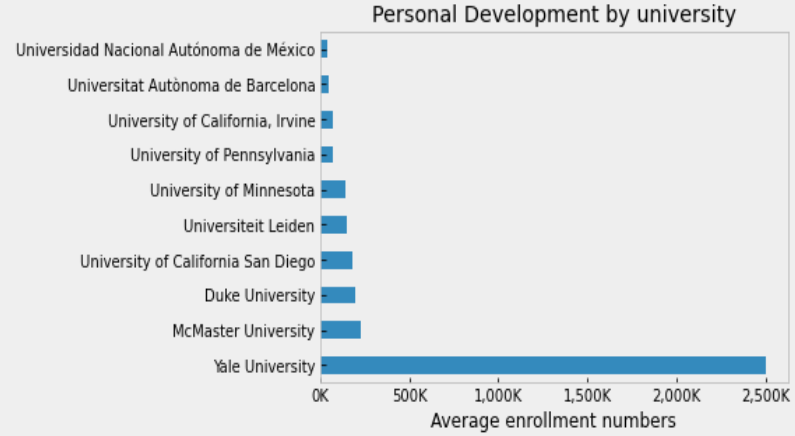
ANOVA

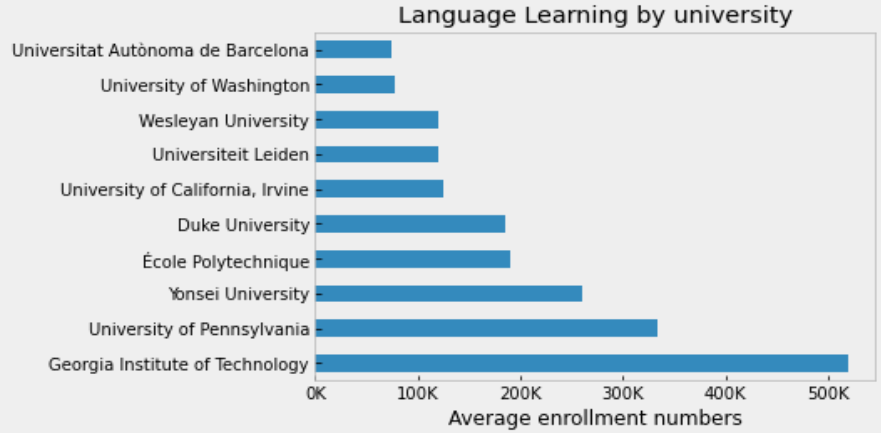
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	453,020,000,000	3	151,007,000,000	**4.6176**	**0.0033**	**2.6149**
Within Groups	29,006,800,000,000	887	32,702,093,316			
Total	29,459,800,000,000	890				

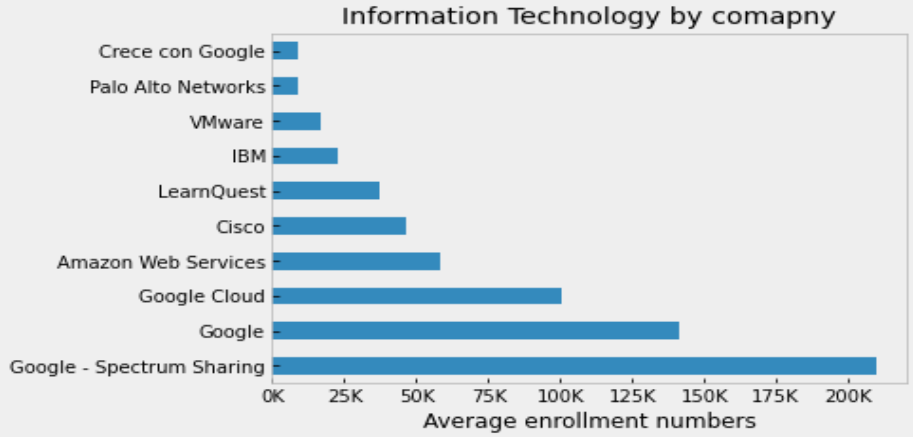
Q#7 – graphs in each category











Regression Results

$$Enrollment_i = \beta_0 + \beta_1 Rating_i + \sum \gamma_i CertificationType_i + \sum \delta_i Difficulty_i + \sum \theta_i Category_i + \sum \rho_i Institution_i + \varepsilon_i$$

OLS Regression Results

Dep. Variable:	num_enrollment	R-squared:	0.218
Model:	OLS	Adj. R-squared:	0.035
Method:	Least Squares	F-statistic:	1.190
Date:	Wed, 07 Jul 2021	Prob (F-statistic):	0.0680
Time:	12:51:46	Log-Likelihood:	-11835.
No. Observations:	890	AIC:	2.401e+04
Df Residuals:	720	BIC:	2.482e+04
Df Model:	169		
Covariance Type:	nonrobust		

course_organization[T.deeplearning.ai]	2.012e+05	1.08e+05	1.871	0.062
-9913.045	4.12e+05			
course_organization[T.École Polytechnique]	1.083e+05	1.63e+05	0.666	0.506
-2.11e+05	4.28e+05			
course_organization[T.École Polytechnique Fédérale de Lausanne]	8.323e+04	8.15e+04	1.021	0.307
-7.68e+04	2.43e+05			
course_organization[T.École des Ponts ParisTech]	-2.829e+04	1.6e+05	-0.177	0.859
-3.42e+05	2.85e+05			
course_rating	1.042e+05	4.18e+04	2.493	0.013
2.22e+04	1.86e+05			