Data Analysis Report: Coursera Course Dataset

Author: Syed Bokhari

Date of Submission: 6th November 2022

Introduction:

The purpose of this project is to analyze the dataset Coursera Course Dataset. The dataset used for this project was acquired from Kaggle. It is a webscraped set created for a Kaggle competition. The dataset comprises of data on hundreds of courses from the Coursera website. Courser is an online learning platform where universities, organizations and content providers offer courses to learners. The dataset and related details can be found at the link: https://www.kaggle.com/datasets/siddharthm1698/coursera-course-dataset?resource=download

The dataset has 7 columns and 891 rows. The columns and their details are listed below.

unnamed column: Contains a unique ID for each course.

course_title: Contains the title of the course, as shown on Coursera platform.

course_organization: The name of the organization offering the course.

course_Certificate_type: The type of certification offered. There are multiple types of certifications offered which are Course, Specialization, Professional Certificate.

course_rating: The average value of ratings given to the course by students who completed it. The ratings are from 5 and lower.

course_difficulty: The level of difficulty of the course. The difficulty levels are: Beginner, Intermediate, Mixed and Advanced.

course_students_enrolled: Lists the total number of students enrolled in the course as provided by Coursera.

Data Exploration:

```
df = pd.read_csv('coursea_data.csv')
```

```
df.head()
```

	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

Fig 1. A snapshot of the dataset columns and some initial values

Initially, the columns of the dataset were investigated to determine the type of data in each column as well as to determine which columns contained categorical variables. The Pandas library in Python was used for this particular purpose. The dataset was saved in a Pandas dataframe. This preliminary investigation revealed that other than the unnamed column containing unique ids which was of int64 type and course_rating column which was of float64 type, all the other columns were of object data type. This slightly complicated things as it meant that the course_students_enrolled column did not contain numerical values which could be compared.

```
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
```

Fig 2. Different Data Types in the Dataset

Next, the categorical columns were identified. It was understood that the unnamed column would not be categorical as it contained unique id's for courses. The course_title column contained a name for each course so had unique values also. Of the remaining five columns, the course_Certificate_type was identified as a categorical column containing three categories: COURSE, SPECIALIZATION and PROFESSIONAL CERTIFICATE. The course_difficulty column also proved to be a categorical column with four categories: Beginner, Intermediate, Mixed and Advanced.

Of the three remaining columns, course_organization had 154 unique values which were too many for it to be considered a categorical column. Interestingly, while the course_rating column was a numerical column, it had 14 unique values only. Finally, the column course_students_enrolled was investigated and found to have 205 unique values. All of these columns were not considered categorical.

A plan was charted for analyzing the data. It was clear that only the course_rating column had numerical values. However, this did not open up many avenues for interesting analyses. A look at the course_students_enrolled column however, revealed that it contained objects of the form '5.3k'. Clearly, these were numbers that could be very useful in analyzing important aspects of the dataset and drawing insights if they could be handled properly. However, this would require some feature engineering and variable transformations.

Data Cleaning and Variable Transformation:

For any meaningful analysis to take place, some data cleaning was required. First, the column names were too unwieldy so they were renamed. The unnamed column was renamed to index, course_title to title, course_organization to organization, course_Certificate_type to certification, course_rating to rating, course_difficulty to difficulty and course_students_enrolled to enrolled. The names were selected for ease of use in subsequent operations. Furthermore, after some consideration, the renamed column index was dropped and the resulting dataframe saved separately as index did not add any value for further analysis.

Next, the enrolled column had to be transformed from an object datatype to a numeric datatype. This was considered necessary because the column had numeric information about students enrolled in a particular course which could be used for several meaningful comparisons in the subsequent analysis. However, the enrolled column saved numbers in a format where 5300 was stored as 5.3k. Furthermore, things were complicated by the fact that some courses had millions of learners and thus had values ending with an m. Therefore, a series of steps were taken to clear non-numeric characters from the enrolled column and save the information conveyed by those characters. For this, the indexes of the rows which had learners in millions were saved. Then, the column was cast as a float type and the information for the characters stored previously was used to properly scale the values to what was originally stored. All values were multiplied by 1000 for k and 1000 again for those indexes where the value stored ended with an m. Ultimately, this resulted in an enrolled column that had numeric values which could be used for some meaningful analyses.

```
df2.head()
```

	title	organization	certification	rating	difficulty	enrolled
0	(ISC)² Systems Security Certified Practitioner...	(ISC)²	SPECIALIZATION	4.7	Beginner	5300.0
1	A Crash Course in Causality: Inferring Causal...	University of Pennsylvania	COURSE	4.7	Intermediate	17000.0
2	A Crash Course in Data Science	Johns Hopkins University	COURSE	4.5	Mixed	130000.0
3	A Law Student's Toolkit	Yale University	COURSE	4.7	Mixed	91000.0
4	A Life of Happiness and Fulfillment	Indian School of Business	COURSE	4.8	Mixed	320000.0

Fig 3. Cleaned and transformed dataset

Exploratory Data Analysis (EDA):

With the cleaned and transformed dataset ready, an EDA could now be conducted. First, the courses with the highest enrollment were noted down. The top 10 most popular courses by enrollment were as follows:

```
df3 = df2.sort_values(by='enrolled', ascending=False)
```

```
df3.head(10)
```

	title	organization	certification	rating	difficulty	enrolled
564	Machine Learning	Stanford University	COURSE	4.9	Mixed	3200000.0
815	The Science of Well-Being	Yale University	COURSE	4.9	Mixed	2500000.0
688	Python for Everybody	University of Michigan	SPECIALIZATION	4.8	Beginner	1500000.0
674	Programming for Everybody (Getting Started wit...	University of Michigan	COURSE	4.8	Mixed	1300000.0
196	Data Science	Johns Hopkins University	SPECIALIZATION	4.5	Beginner	830000.0
129	Career Success	University of California, Irvine	SPECIALIZATION	4.4	Beginner	790000.0
261	English for Career Development	University of Pennsylvania	COURSE	4.8	Mixed	760000.0
765	Successful Negotiation: Essential Strategies a...	University of Michigan	COURSE	4.8	Mixed	750000.0
199	Data Science: Foundations using R	Johns Hopkins University	SPECIALIZATION	4.6	Beginner	740000.0
211	Deep Learning	deeplearning.ai	SPECIALIZATION	4.8	Intermediate	690000.0

Fig 4. Top 10 courses by student enrollment

Next, the top 12 courses by rating were noted. As many courses could have identical ratings, the enrolled column was chosen as the tie breaker.

```
df4 = df2.sort_values(by=['rating','enrolled'], ascending=[False,False])
```

```
df4.head(12)
```

	title	organization	certification	rating	difficulty	enrolled
432	Infectious Disease Modelling	Imperial College London	SPECIALIZATION	5.0	Intermediate	1600.0
251	El Abogado del Futuro: Legaltech y la Transfor...	Universidad Austral	COURSE	5.0	Beginner	1500.0
564	Machine Learning	Stanford University	COURSE	4.9	Mixed	3200000.0
815	The Science of Well-Being	Yale University	COURSE	4.9	Mixed	2500000.0
626	Neural Networks and Deep Learning	deeplearning.ai	COURSE	4.9	Intermediate	630000.0
684	Python Data Structures	University of Michigan	COURSE	4.9	Mixed	420000.0
322	First Step Korean	Yonsei University	COURSE	4.9	Beginner	400000.0
427	Improving Deep Neural Networks: Hyperparameter...	deeplearning.ai	COURSE	4.9	Beginner	270000.0
512	Introduction to Psychology	Yale University	COURSE	4.9	Beginner	270000.0
162	Convolutional Neural Networks	deeplearning.ai	COURSE	4.9	Intermediate	240000.0
291	Excel Skills for Business	Macquarie University	SPECIALIZATION	4.9	Beginner	240000.0
124	COVID-19 Contact Tracing	Johns Hopkins University	COURSE	4.9	Beginner	220000.0

Fig 5. Top 12 courses by ratings

Furthermore, plots were constructed to determine what kind of relationships were present in the dataset. The first such plot looked at the relationship between course rating and the number of students enrolled in the course:

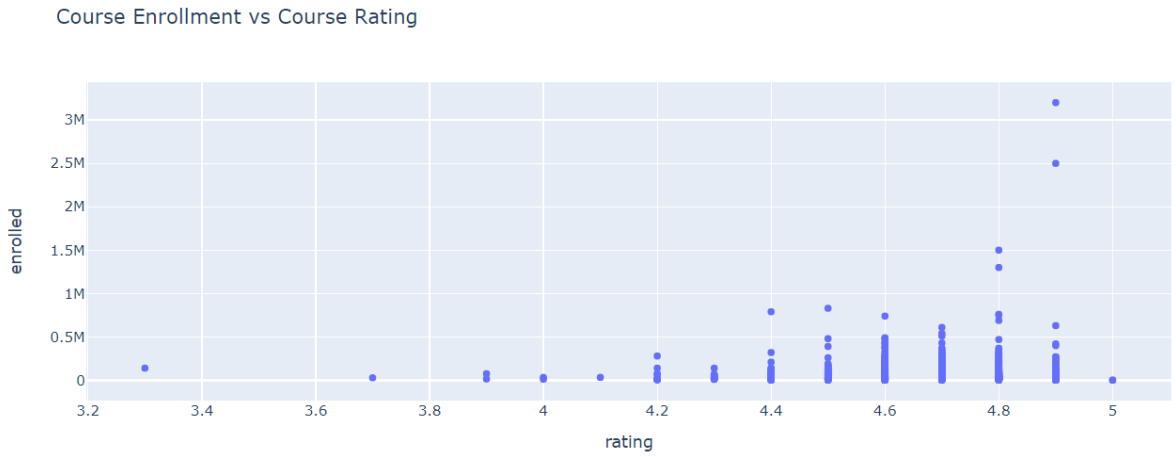

Fig 6. Scatter plot of Course Enrollment vs. Course Rating

From the plot it can be seen that courses with lower enrollments have lower enrollments and vice versa. There seems to be a relationship between the two. To see this relationship a little more clearly, a bar plot was constructed:

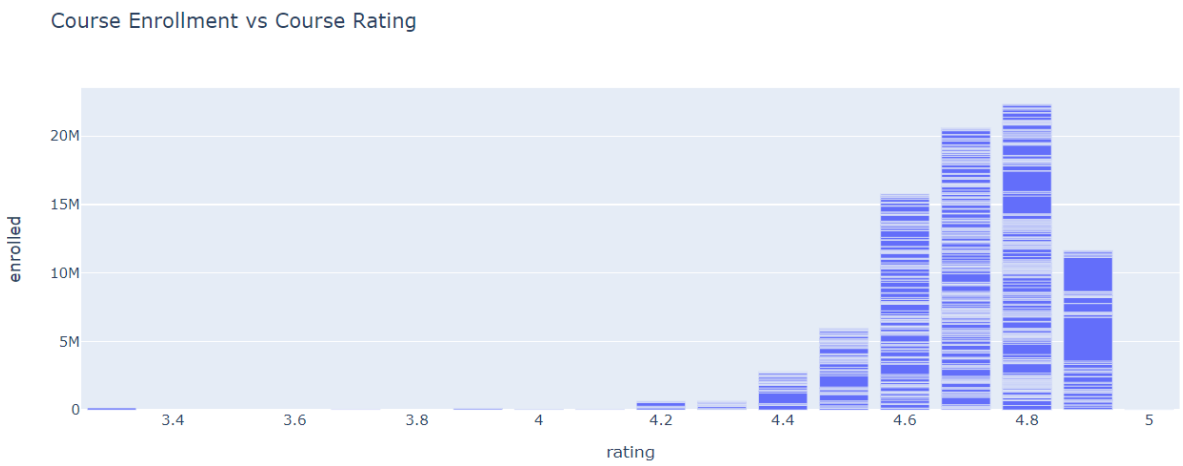

Fig.7 Bar plot of Course Enrollment vs. Course Rating

Now, the relationship between the two is much more clearly defined. We see that there seems to be an almost exponential increase in enrollments for courses as ratings improve till 4.8. Afterwards however, there is a decline for 4.9. Given the nature of the ratings system, it is very difficult for any course to be consistently rated 5 so lack of values there is understandable.

Next, the number of enrollments was compared to the certification level of the course.

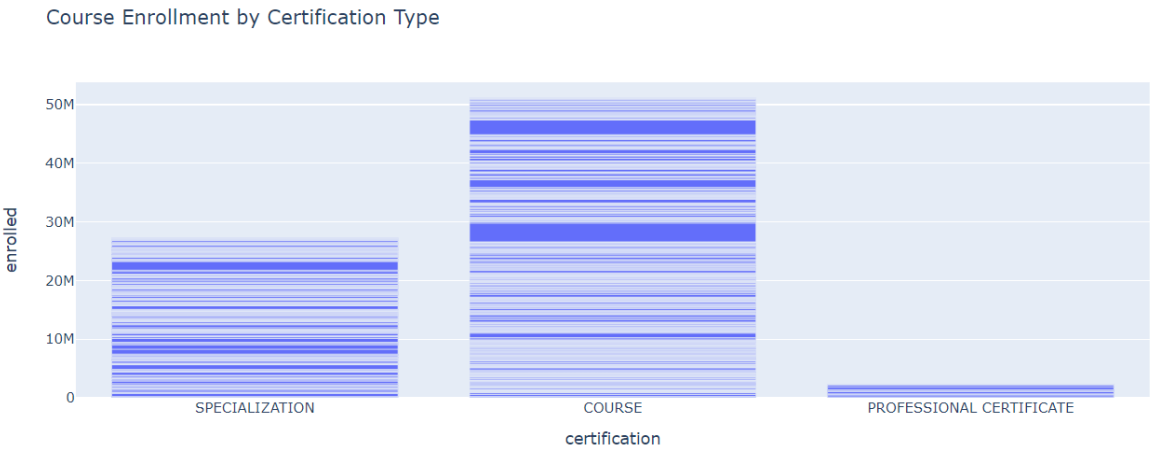

Fig 8. Bar plot of certification type and enrollment

The number of enrollments is highest for individual Courses and much lesser for Specializations. It is very low for Professional Certificates. For context, a Specialization consists of multiple courses tied to a topic. A Professional Certificate is even more rigorous, sometimes spanning two or more specializations. Thus, these results are as expected.

Subsequently, course enrollment was plotted against course difficulty:

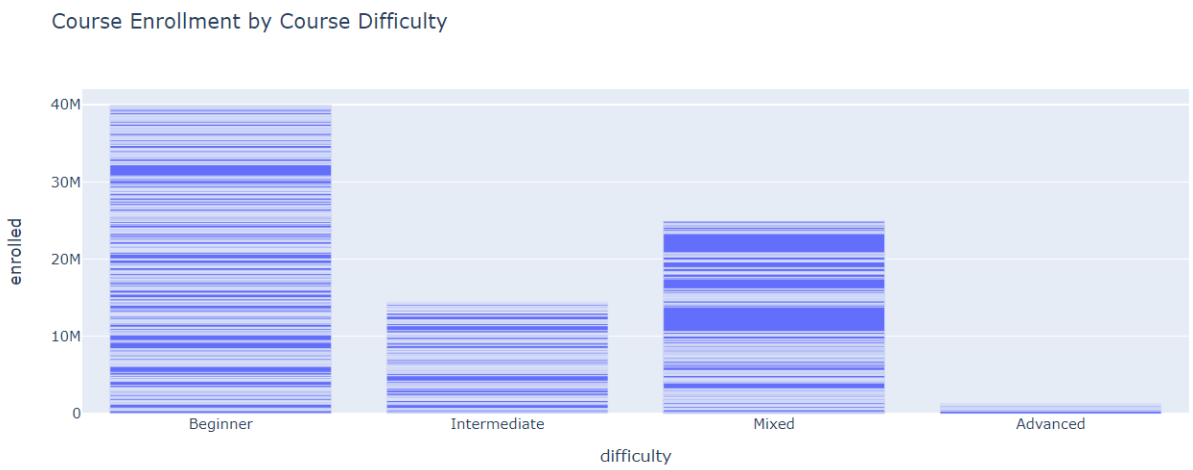

Fig 9. Bar plot of course difficulty and enrollment

The plot shows the highest enrollments for Beginner level courses, followed by Mixed level, then Intermediate level and finally Advanced level. For context, Beginner level courses require little prior knowledge, Mixed level courses require some prior knowledge but not too much, Intermediate level courses require prior knowledge and skills in the area and Advanced level courses require substantial knowledge and skill in the area. Thus, once more, these results follow the trend that is expected.

Afterwards, a box plot of course difficulty and course rating was plotted:

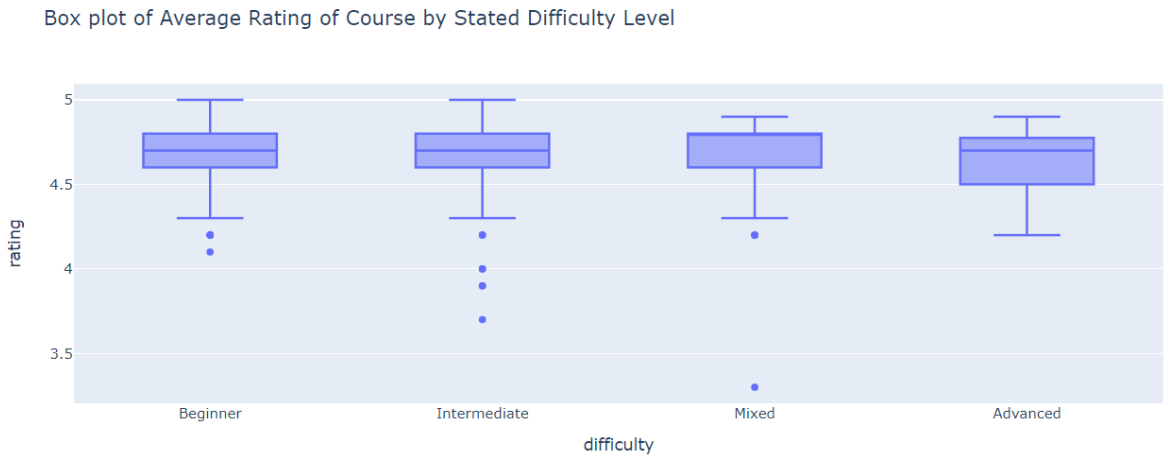

Fig 10. Box plot of course rating for different course difficulty levels

There appears to be a slight trend of lower ratings for more difficult courses but overall things do not appear to be clear. Mixed courses are supposed to be easier than Intermediate courses but they have a lower rating score than Intermediate courses.

Finally, a box plot of course certification type and course rating was plotted:

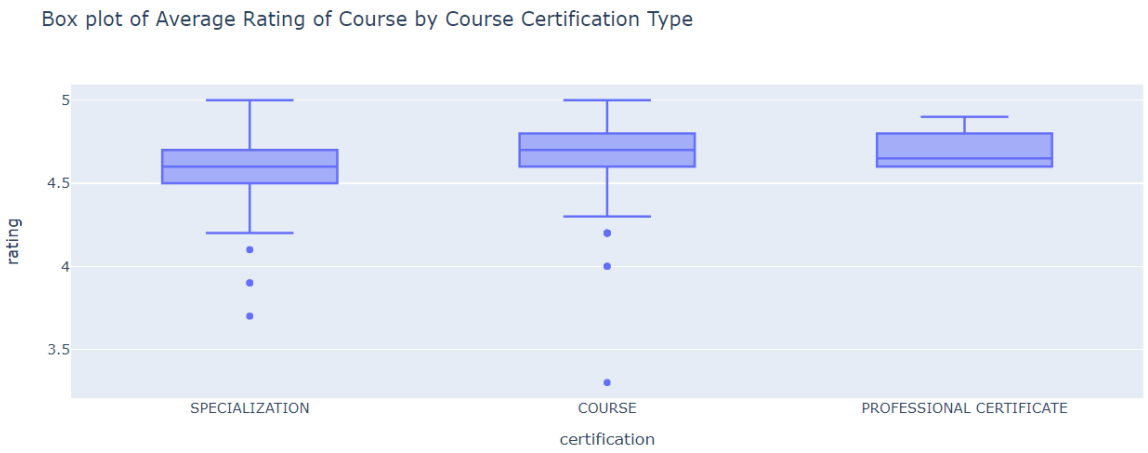

Fig 11. Box plot of course rating for different course difficulty levels

Once more, there appears to be no clarity on what trend if any is there. Professional Certificates tend to have a higher rating score in general than Specializations.

Key Insights from EDA

Some key insights from the EDA were derived which were as follows:

1. There is clearly a relationship between course enrollments and course ratings. The EDA showed results that appeared close to a normal distribution that was skewed to the left.
2. Enrollments in courses tend to decrease with more advanced certification types.
3. Enrollments in courses tend to decrease with more difficult course levels.
4. There appears to be no clear relationship between course rating and course difficulty.
5. There appears to be no clear relationship between course rating and course certification type.

Key Questions

The insights derived from EDA lead to several questions. These questions were converted into Null and Alternative Hypotheses for further testing using statistical tests of significance:

1. Is there a relationship between course difficulty level and course ratings?
2. Is there a relationship between course certification type and course ratings?
3. Is there a relationship between course difficulty level and number of enrollments?
4. Is there a relationship between course certification type and number of enrollments?

And finally an additional question may be added to these:

5. Is there a relationship between course certification type and course difficulty level?

Hypothesis Formulation:

Hypothesis 1

For the first key question, it is not clear whether a course's stated difficulty level has an impact on the ratings it gets. A higher level of difficulty should drag down a course's ratings. On the other hand, only true enthusiasts would want to take a higher difficulty level course and they may thus give such courses higher ratings.

The first hypothesis can be formulated out of this question as the following Null and Alternative Hypotheses:

H01: μ1= μ2= μ3= μ4 The mean ratings of courses with Beginner, Mixed, Intermediate and Advanced difficulty is the same.

HA1: At least one of the mean ratings for the courses with different difficulty levels is not the same.

Hypothesis 2

For the second key question, the reasonable answer might be yes. But to be certain of this, it can also be formulated as a hypothesis.

The second hypothesis can be formulated as the following Null and Alternative Hypotheses:

H02: μ1= μ2= μ3 The mean ratings of courses with Course, Specialization and Professional Certificate certification type is the same.

HA2: At least one of the mean ratings for courses with different certification types is not the same.

Hypothesis 3

For the third key question, the answer should be yes. But it too is formulated as a hypothesis nevertheless.

The third hypothesis can be formulated as the following Null and Alternative Hypotheses:

H03: μ1= μ2= μ3= μ4 The mean enrollment of courses with Beginner, Mixed, Intermediate and Advanced difficulty is the same.

HA3: At least one of the mean enrollments for courses with different difficulty levels is not the same.

Hypothesis 4

For the fourth key question, the hypothesis formulated is as follows:

H04: μ1= μ2= μ3 The mean enrollment of courses with Course, Specialization and Professional Certificate certification type is the same.

HA4: At least one of the mean enrollments for courses with different certification types is not the same.

Hypothesis 5

For the final key question, the hypothesis formulated is as follows:

H05: Course certification and course difficulty have no relationship.

HA5: Course certification and course difficulty do have a relationship.

Hypothesis Testing:

Now the various hypotheses are tested using tests of significance. The significance level for the p-value is set to 0.95. So, the alpha level is 0.05 for all hypotheses.

The ANOVA test was conducted using the statsmodel package in Python. The appropriate formula for each ANOVA test along with the requisite data is fitted to an object which is passed to the ANOVA object.

The details of the Chi-Square test are stated where it was used.

Next, we test each hypothesis with the appropriate test of significance.

Hypothesis 1

The appropriate test for this is the ANOVA test. The results are as follows:

	df	sum_sq	mean_sq	F	PR(>F)
difficulty	3.0	0.490805	0.163602	6.32825	0.000301
Residual	887.0	22.931238	0.025853	NaN	NaN

Fig 12. ANOVA results for Hypothesis 1

The p-value is lesser than the alpha value of 0.05. Therefore, the null hypothesis is rejected and the conclusion is that the mean ratings for courses with different difficulty levels are not all the same.

Hypothesis 2

The appropriate test for this is the ANOVA test. The results are as follows:

	df	sum_sq	mean_sq	F	PR(>F)
certification	2.0	1.559108	0.779554	31.662894	5.213883e-14
Residual	888.0	21.862935	0.024620	NaN	NaN

Fig 13. ANOVA results for Hypothesis 2

The p-value is lesser than the alpha value of 0.05. Therefore, the null hypothesis is rejected and the conclusion is that the mean ratings for courses with different certification types is not the same.

Hypothesis 3

The appropriate test for this is the ANOVA test. The results are as follows:

	df	sum_sq	mean_sq	F	PR(>F)
difficulty	3.0	4.530204e+11	1.510068e+11	4.617649	0.00326
Residual	887.0	2.900676e+13	3.270209e+10	NaN	NaN

Fig 14. ANOVA results for Hypothesis 3

The p-value is lesser than the alpha value of 0.05. Therefore, the null hypothesis is rejected and the conclusion is that the mean enrollments for courses with different difficulty levels are not all the same.

Hypothesis 4

The appropriate test for this is the ANOVA test. The results are as follows:

	df	sum_sq	mean_sq	F	PR(>F)
certification	2.0	1.250471e+11	6.252355e+10	1.892668	0.151277
Residual	888.0	2.933473e+13	3.303461e+10	NaN	NaN

Fig 15. ANOVA results for Hypothesis 4

The p-value is greater than the significance level 0.05. Therefore, the test fails to reject the null hypothesis. Interestingly enough, this indicates that the mean enrollments for courses are not affected by the type of certification!

Hypothesis 5

The appropriate test for this hypothesis is the Chi-square test.

First, a contingency table for the two categorical variables is prepared. Then the relationship between the two variables is plotted.

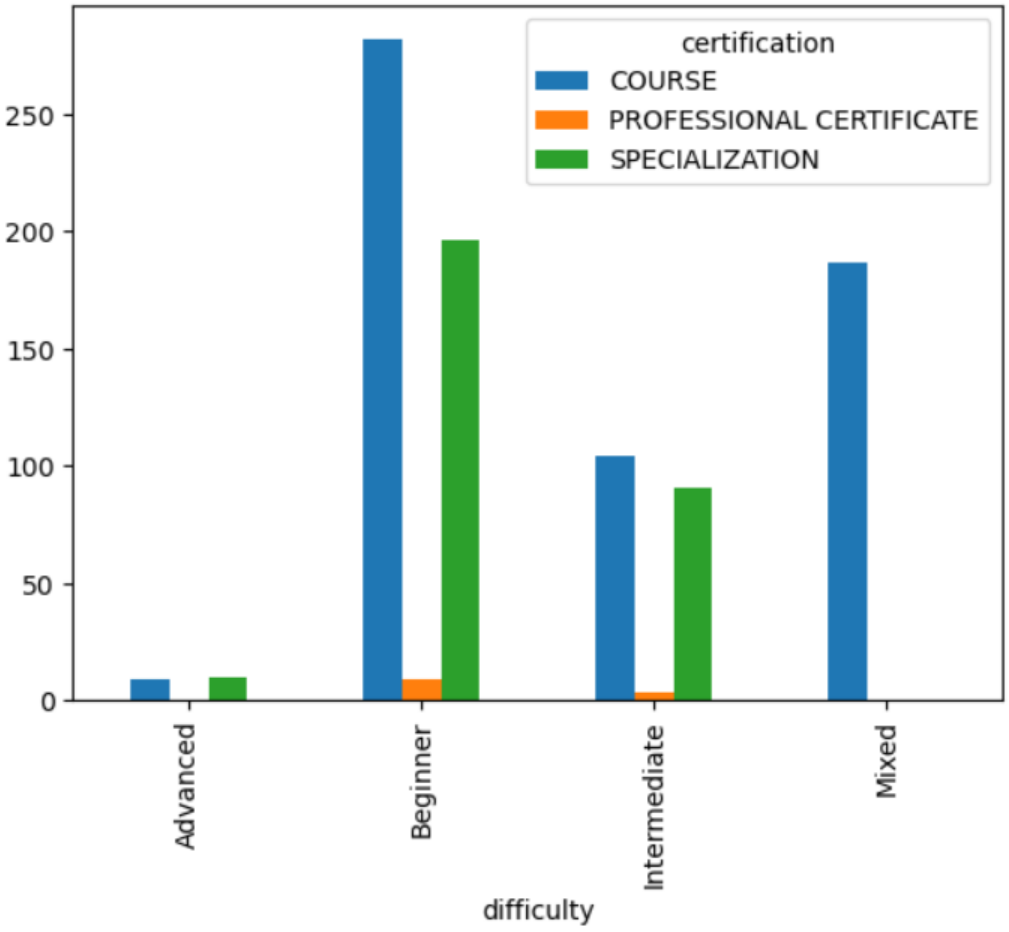

Fig 16. Contingency plot

Next, the Chi-square test is conducted. The results are as follows:

```
chi-square statistic: 129.06235917120085 , p_value: 2.025032001515206e-25 , degree of freedom: 6 ,expected frequencies: [[1.24107744e+01 2.55892256e-0
1 6.33333333e+00]
 [3.18107744e+02 6.55892256e+00 1.62333333e+02]
 [1.29333333e+02 2.66666667e+00 6.60000000e+01]
 [1.22148148e+02 2.51851852e+00 6.23333333e+01]]
```

The p-value is lesser than the alpha value of 0.05. Therefore, the null hypothesis is rejected and the conclusion is that the course difficulty is affected by course certificate type.

Hypotheses Results:

The hypothesis testing revealed the following answers to the key questions:

1. The mean ratings for courses with different difficulty levels are not all the same.
2. The mean ratings for courses with different certification types is not the same.
3. The mean enrollments for courses with different difficulty levels are not all the same.
4. Mean enrollments for courses are not affected by the type of certification.
5. Course difficulty is affected by course certificate type

Advanced Analysis:

Although hypothesis testing is a good way to gain insights, advanced analysis for this dataset is possible. The best tool for this would be regression analysis. With regression analysis, relationships between the variables could be quantified. An additional analysis should be conducted between course enrollments and course ratings. The preliminary EDA revealed that the enrollment rates seemed to be distributed according to the normal distribution across course ratings with a leftward skew in the distribution. The data could be standardized and investigated further for more interesting insights.

Dataset Quality:

The dataset used for the analysis was of decent quality. Although relatively small at only 891 rows and 7 columns, it was nevertheless a dataset of good quality. There were no missing values. However, the dataset was too limited. Additional attributes could have made it richer ad resulted in more insights. For instance, how long had a course been made available when the data was collected? A course with several million enrollments looks impressive at first glance. However, it may be one of the first courses made available on the platform which would allow it to accumulate so many enrollments over a period of several years. In contrast, a course with several hundred thousand enrollments over one year may have more future potential. This would allow for more interesting variables to be defined. Another important attribute would be course category. This could allow for more interesting comparisons and insights on which categories are more popular than others etc.

Appendix

Data Exploration:

```
df['course_difficulty'].value_counts()
```

```
Beginner        487
Intermediate    198
Mixed           187
Advanced         19
Name: course_difficulty, dtype: int64
```

```
df['course_Certificate_type'].value_counts()
```

```
COURSE                      582
SPECIALIZATION              297
PROFESSIONAL CERTIFICATE     12
Name: course_Certificate_type, dtype: int64
```

```
df['course_difficulty'].value_counts()
```

```
Beginner        487
Intermediate    198
Mixed           187
Advanced         19
Name: course_difficulty, dtype: int64
```

```
df['course_organization'].value_counts()
```

```
University of Pennsylvania                    59
University of Michigan                        41
Google Cloud                                  34
Johns Hopkins University                      28
Duke University                               28
                                              ..
Nanyang Technological University, Singapore    1
ScrumTrek                                      1
JetBrains                                      1
Tsinghua University                            1
Mail.Ru Group                                  1
Name: course_organization, Length: 154, dtype: int64
```

```
df['course_rating'].value_counts()
```

```
4.8    256
4.7    251
4.6    168
4.5     80
4.9     68
4.4     34
4.3     15
4.2     10
5.0      2
4.0      2
3.9      2
3.3      1
4.1      1
3.7      1
Name: course_rating, dtype: int64
```

Feature Engineering:

```
df2=df1
```

```
mils=[]
for i, value in enumerate(df2['enrolled']):
    if value.endswith('k'):
        df2['enrolled'][i] = value.replace('k','')
    if value.endswith('m'):
        mils.append(i)
        df2['enrolled'][i] = value.replace('m','')
```

```
df2['enrolled'] = df2['enrolled'].astype('float64')
```

```
df2['enrolled'] = df2['enrolled']*1000
```

```
df2['enrolled'][mils] = df2['enrolled'][mils]*1000
```

```
df2.head()
```

	title	organization	certification	rating	difficulty	enrolled
0	(ISC)² Systems Security Certified Practitioner...	(ISC)²	SPECIALIZATION	4.7	Beginner	5300.0
1	A Crash Course in Causality: Inferring Causal...	University of Pennsylvania	COURSE	4.7	Intermediate	17000.0
2	A Crash Course in Data Science	Johns Hopkins University	COURSE	4.5	Mixed	130000.0
3	A Law Student's Toolkit	Yale University	COURSE	4.7	Mixed	91000.0
4	A Life of Happiness and Fulfillment	Indian School of Business	COURSE	4.8	Mixed	320000.0

EDA (Most popular organization by enrollment):

```
df6 = df2.groupby('organization')['enrolled'].sum().sort_values(ascending=False)
```

```
df6.head(20)
```

```
organization
University of Michigan                      7437700.0
University of Pennsylvania                  5501300.0
Stanford University                         4854000.0
University of California, Irvine            4326000.0
Johns Hopkins University                    4298900.0
Duke University                             3967600.0
Yale University                             3952000.0
IBM                                         2956400.0
deeplearning.ai                             2863400.0
Google Cloud                                2604300.0
Georgia Institute of Technology             1813000.0
University of Illinois at Urbana-Champaign  1679000.0
University of Virginia                      1556000.0
Berklee College of Music                    1288000.0
University of California, Davis             1278700.0
Universidad Nacional Autónoma de México     1207900.0
University of California San Diego          1166000.0
Google                                      1153000.0
The University of Edinburgh                  875000.0
University of Washington                     837000.0
Name: enrolled, dtype: float64
```

EDA (Most popular organization by rating):

```
df7 = df2.groupby('organization')['rating'].mean().sort_values(ascending=False)
```

```
df7.head(10)
```

```
organization
Hebrew University of Jerusalem                4.900000
Nanyang Technological University, Singapore   4.900000
Universidade Estadual de Campinas             4.900000
Crece con Google                              4.900000
London Business School                        4.900000
Google - Spectrum Sharing                     4.900000
ScrumTrek                                     4.900000
Universidade de São Paulo                     4.866667
The University of Chicago                     4.850000
Universidad de los Andes                      4.820000
Name: rating, dtype: float64
```

ANOVA Testing (Hypothesis 1):

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
```

```
formula ='rating ~ difficulty'
model=ols(formula, df2).fit()
```

```
aov_table = anova_lm(model)
aov_table
```

	df	sum_sq	mean_sq	F	PR(>F)
difficulty	3.0	0.490805	0.163602	6.32825	0.000301
Residual	887.0	22.931238	0.025853	NaN	NaN

ANOVA Testing (Hypothesis 2):

```
formula4 ='rating ~ certification'
model4=ols(formula4, df2).fit()
```

```
aov_table4 = anova_lm(model4)
aov_table4
```

	df	sum_sq	mean_sq	F	PR(>F)
certification	2.0	1.559108	0.779554	31.662894	5.213883e-14
Residual	888.0	21.862935	0.024620	NaN	NaN

ANOVA Testing (Hypothesis 3):

```
formula2 ='enrolled ~ difficulty'
model2=ols(formula2, df2).fit()
```

```
aov_table2 = anova_lm(model2)
aov_table2
```

	df	sum_sq	mean_sq	F	PR(>F)
difficulty	3.0	4.530204e+11	1.510068e+11	4.617649	0.00326
Residual	887.0	2.900676e+13	3.270209e+10	NaN	NaN

ANOVA Testing (Hypothesis 4):

```
formula3 ='enrolled ~ certification'
model3=ols(formula3, df2).fit()
```

```
aov_table3 = anova_lm(model3)
aov_table3
```

	df	sum_sq	mean_sq	F	PR(>F)
certification	2.0	1.250471e+11	6.252355e+10	1.892668	0.151277
Residual	888.0	2.933473e+13	3.303461e+10	NaN	NaN

Chi-Square Testing (Hypothesis 5):

```
contingency = pd.crosstab(df2['difficulty'], df2['certification'])
contingency
```

certification	COURSE	PROFESSIONAL CERTIFICATE	SPECIALIZATION
difficulty			
Advanced	9	0	10
Beginner	282	9	196
Intermediate	104	3	91
Mixed	187	0	0

```
chi2, p_val, dof, exp_freq = chi2_contingency(contingency, correction = False)
print('chi-square statistic: {} , p_value: {} , degree of freedom: {} ,expected frequencies: {}'.format(chi2, p_val, dof, exp_freq))
```

```
chi-square statistic: 129.06235917120085 , p_value: 2.025032001515206e-25 , degree of freedom: 6 ,expected frequencies: [[1.24107744e+01 2.55892256e-0
1 6.33333333e+00]
 [3.18107744e+02 6.55892256e+00 1.62333333e+02]
 [1.29333333e+02 2.66666667e+00 6.60000000e+01]
 [1.22148148e+02 2.51851852e+00 6.23333333e+01]]
```