Institute of Technology, Nirma University

Sessional Assignment - Report

Course Code and Name	:	2CEOE76 - Scientific Programming
Name of the Student	:	Sanjaykumar J. Parmar
Roll Number	:	18BEC069
Topic of Your Report	:	Coursera Database Analysis
Date of Submission	:	6th December, 2020
Submitted To	:	Dr. Jaiprakash Verma

Table of Contents

1. Information

Coursera is an American *massive open online course (MOOC)* provider founded in 2012 by **Stanford University's** computer science **professors Andrew Ng and Daphne Koller** that offers massive open online courses (MOOC), specializations, degrees, professional and mastertrack courses.

Coursera works with universities and other organizations to offer online courses, certifications, and degrees in a variety of subjects. Since its launch, this site has served more than 25 million students in more than 2,000 classes.

As of December 2019, the total number of partners is more than 200 across 29 countries. Coursera mainly works with universities and colleges, but also with corporations and governments. University partners include *University of São Paulo in Brazil*, *University of London in the UK*, *Indian School of Business of India*, *Yonsei University in Korea*, and institutions like *Yale*, *University of Illinois* and *University of Pennsylvania*.

Coursera courses last approximately four to twelve weeks, with one to two hours of video lectures a week. These courses provide *quizzes*, *weekly exercises*, *peer-graded and reviewed assignments*, *an optional Honors assignment* and sometimes *a final project* or *exam* to complete the course. Courses are also provided on-demand, in which case users can take their time in completing the course with all of the material available at once. As of May 2015, Coursera offered 104 on-demand courses it also provides guided projects which are short 2-3 hour projects that can be done.

2. Objective

In this project I will be going to make the Coursera Course Analysis. This aims the new learners to get the right course to learn by just answering few questions. On the basis of your area of interest we will be suggesting you the courses with the details like, Number of students Enrolled, University Name, Duration of Course and most important the overall rating of the course. It will create the **Intelligent Course Recommendation System**. Hence we had to scrap some data from some educational websites. Here I have scraped data from Coursera Website.

3. Scope

This intelligent course recommendation system will be helping the students in selecting the courses in their are of interest from the variety of Universities, and . It will also be helpful to the other universities like, they will get an overall idea about what is the demand of industry and students, also with the comparison of courses of other institutes they will get the idea about the area of improvements.

4. Dataset Selection

The dataset which I have used for the Analysis is been created by Mr. Siddharth, He had created the dataset by scraping the data from official website of coursera. This dataset contains the information regarding Course Title, Organization, Certification Type, Rating, Difficulty Level, Number of Students Enrolled.

5. Characteristics of Dataset

This dataset contains mainly 6 columns and 890 course data.
This is the detailed description :

I. **Title :** Contains the course title.

Out of the 890 Course Titles the 888 courses are unique.

II. **Organization :** It tells which organization is conducting the courses.

This 890 Courses have been offered by 154 different Organizations.

III. **Certificate Type :** It has details about what are the different certifications available in courses.

There are tree different types of certificate type : Course, Specialization and other.

i. Course : 65%
ii. Specialization : 33%
iii. Others : 1%

IV. **Rating :** It has the ratings associated with each course.

On the basis of data of 890 courses.

i. Average Rating : 4.68
ii. Std. Deviation : 0.16
iii. Minimum : 3.3
iv. Maximum : 5

V. **Difficulty Level :** It tells about how difficult or what is the level of the course.

i. Beginner : 55%
ii. Intermediate : 22%
iii. Others : 23%

VI. **Students Enrolled :** It has the number of students that are enrolled in the course.

6. Methodology and Concepts

First of all with the help of Pandas Library, I have loaded the Dataset(coursera_data.csv) into the jupiter notebook.

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

Then viewed the column's information[*Output 1*],

```
df.columns
```

```
Index(['Unnamed: 0', 'course_title', 'course_organization',
       'course_Certificate_type', 'course_rating', 'course_difficulty',
       'course_students_enrolled'],
      dtype='object')
```

Output 1

As the dataset contains 891 rows x 7 columns, here, [*Output 2*] shows the first 9 rows of the dataset using **`df[0:9]`**.

```
df[0:9]
```

	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
5	661	ADHD: Everyday Strategies for Elementary Students	University at Buffalo	COURSE	4.7	Beginner	39k
6	54	AI For Everyone	deeplearning.ai	COURSE	4.8	Beginner	350k
7	488	AI For Medical Treatment	deeplearning.ai	COURSE	4.8	Intermediate	2.4k
8	58	AI Foundations for Everyone	IBM	SPECIALIZATION	4.7	Beginner	61k

Output 2

As the number of students Enrolled in K and M, here [*Output 3*] I have converted it to the 1000s and 1000000s.

```
df['course_students_enrolled']=df['course_students_enrolled'].str.replace('k', '*1000')
```

```
df['course_students_enrolled']=df['course_students_enrolled'].str.replace('m', '*1000000')
```

```
df['course_students_enrolled'] = df['course_students_enrolled'].map(lambda x: eval(x))
```

Output 3

After converting it to the 1000s and 1000000s, lets plot the data of students enrolled in each course [*Output 4*].

```
0        5300.0
1       17000.0
2      130000.0
3       91000.0
4      320000.0
         ...
886     52000.0
887     21000.0
888     30000.0
889      9800.0
890     38000.0
Name: course_students_enrolled, Length: 891, dtype: float64
```

Output 4

The following image[*Output 5*] indicates the statistical analysis of overall courses offered, their ratings and number of students enrolled.

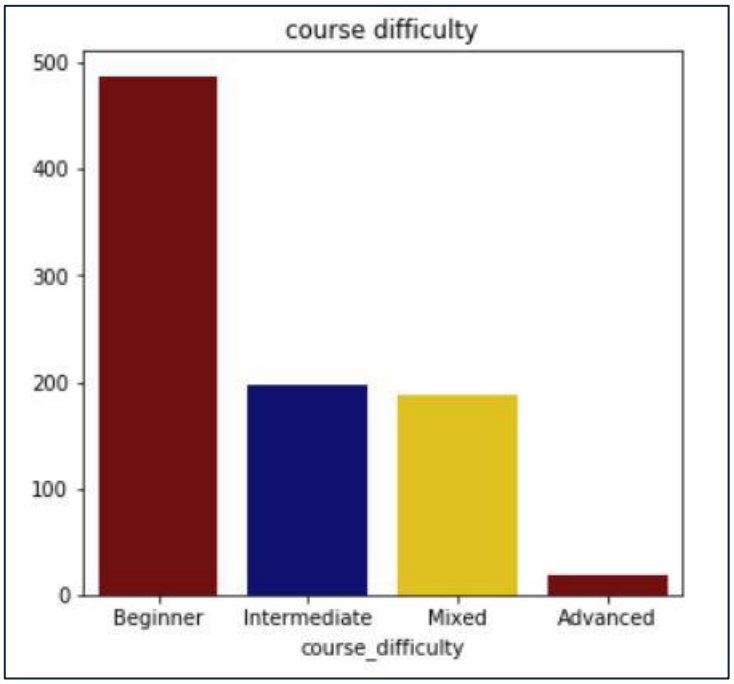
```
df.describe()
```

	Unnamed: 0	course_rating	course_students_enrolled
count	891.000000	891.000000	8.910000e+02
mean	445.000000	4.677329	9.055208e+04
std	257.353842	0.162225	1.819365e+05
min	0.000000	3.300000	1.500000e+03
25%	222.500000	4.600000	1.750000e+04
50%	445.000000	4.700000	4.200000e+04
75%	667.500000	4.800000	9.950000e+04
max	890.000000	5.000000	3.200000e+06

Output 5

Output 6

As the database contains the information related to the course certification Type, so on the basis of that data, here [*Output 6*] I have plotted the bar graph of Course Certification Types. And from which anyone can conclude that, nearly more than 550+ courses are their out of data of 890, and there are very few number of Professional Certifications.



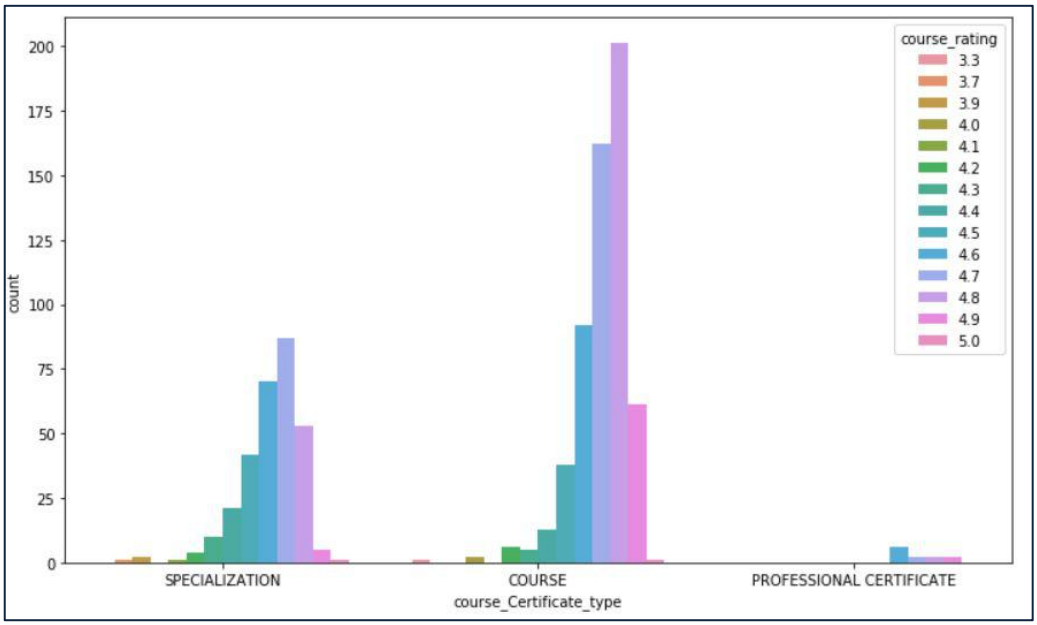
Output 7

The above figure [*Output 8*] indicates the bar graph of Difficulty Level of Courses available in the dataset or you can say Coursera. And from it I can conclude that, coursera is very good platform for gaining the basic(i.e. beginner level) knowledge in different fields as nearly 480 courses are of Beginner Level. And there are very few i.e. only 10-15 courses are there with the advanced level. So, after gaining the basics of topic you have to find some other Educational sites for the advanced learning of that things.
Both of the above graphs(i.e. *Output 6* & *Output 7*) have been plotted using the code given below.

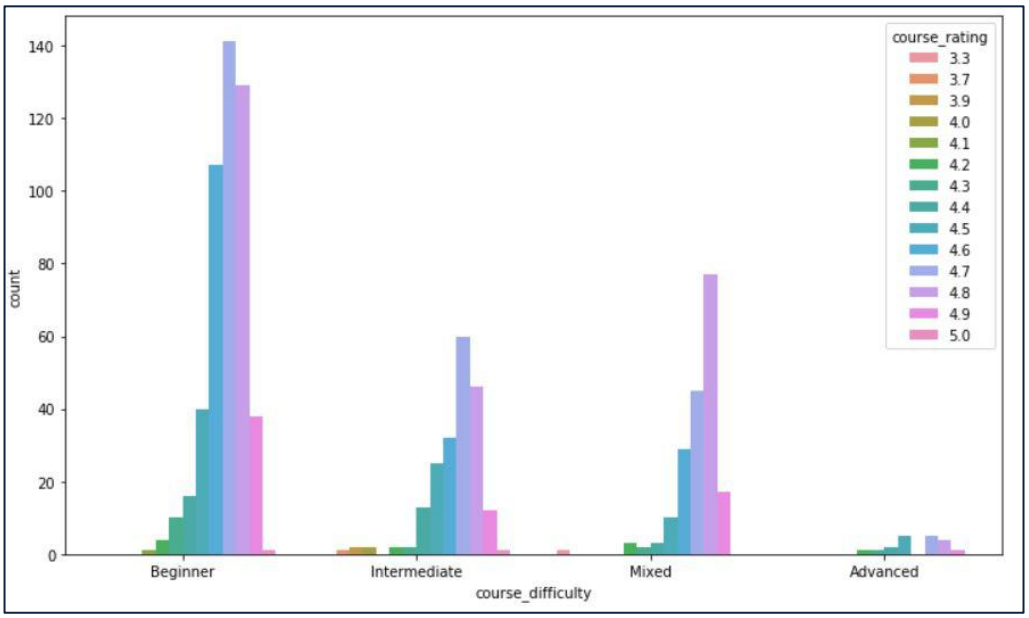
```
f,ax=plt.subplots(1,2,figsize=(15,5))
colours = ["maroon", "navy", "gold"]
sns.countplot('course_Certificate_type',ax=ax[0],data=df,palette=colours)
ax[0].set_title('course Certificate type')
ax[0].set_ylabel('')
sns.countplot('course_difficulty',ax=ax[1],data=df,palette=colours)
ax[1].set_title('course difficulty')
ax[1].set_ylabel('')
plt.subplots_adjust(wspace=0.8)
plt.show()
```

Code 1

As in the above plots we have seen the graphical information related to that of Course Certificate Type and Course Difficulty [*Output 6 & Output 7*], lets have a look at different types of Certification Type and Difficulty Level w.r.t that of Rating of that particular course.



Output 8 : Course Certificate Time w.r.t Course Rating



Output 9 : Difficulty Level w.r.t Course Rating

The above graph have been plotted using Seaborn Library, to have course certificate type graph w.r.t that of course rating using command [***sns.countplot( )***].

```
fig=plt.figure(figsize=(10,6))

sns.countplot('course_Certificate_type',data=df,hue='course_rating' )
plt.tight_layout()
plt.show()

fig=plt.figure(figsize=(10,6))
sns.countplot('course_difficulty',data=df,hue='course_rating' )
plt.tight_layout()
plt.show()
```

After the analysis of database through Certificate Type and Difficulty Level, Lets have a look at Course Organization(i.e The institute which is providing the course).
Here, [*Output 10*] using the keyword ***groupby, rename and sort_values & reset_index*** I have created the new column containing unique Course Organization/ Institute with the number of courses offered by that Organization/ Institute. The name of organizations have been sorted on the basis of number of courses offered in descending order. In this [*Output 10*], I have shown the Top 10 Institutes by number of Course offered, the code for the same is given below[*code 2*].

	course_organization	Count
0	University of Pennsylvania	59
1	University of Michigan	41
2	Google Cloud	34
3	Duke University	28
4	Johns Hopkins University	28
5	University of California, Irvine	27
6	University of Illinois at Urbana-Champaign	22
7	IBM	22
8	University of California, Davis	21
9	University of Colorado Boulder	19

Output 10

From the above list of Top 10 Universities/ Institutes, it is confirm that University of Pennesylvania is leading the chart of highest number of courses offered[59], followed by University of Michigan[41] and Google Cloud[34].

```
ct=df.groupby(['course_title'])['course_title'].count()

co=df.groupby(['course_organization'])['course_organization'].count()

co=pd.DataFrame(co)
co=co.rename(columns={"course_organization": "Count"})
co=co.sort_values(by=['Count'], ascending=False)
co=co.reset_index()
co_top_10=co.head(10)
co_top_10
```

Code 2

Similar to that of above, Lets have a let's have a courses name arranged in the descending order on the basis of number of students enrolled. From this we can get the idea about what is the demand of students and on that basis the Universities/Institutes may get the idea about which courses they could offer in future which will get attracted by students and can get more enrollments. Here, [*Output 11*] shows that ***3.2M*** students have enrolled for the courses on machine learning which is highest and followed by Science of well-Being[2.5M] and Python[1.5M]. From this we can conclude that the hot topics in the market are *Machine Learning, Programming and Data Science*. And from that we can also say that the industry demands it, because in the era where everything has became digital it requires this three things only.

The code to get below data i.e. Top 10 Courses on the basis of Students enrolled is [*Code 3*], in which we just need to sort the database in descending order.

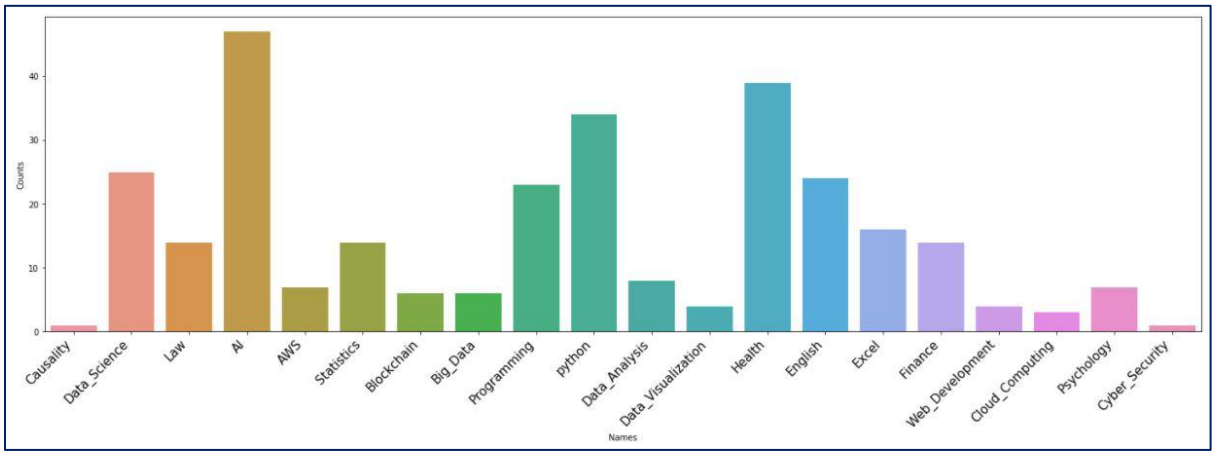
	course_title	course_students_enrolled
564	Machine Learning	3200000.0
815	The Science of Well-Being	2500000.0
688	Python for Everybody	1500000.0
674	Programming for Everybody (Getting Started wit...	1300000.0
196	Data Science	830000.0
129	Career Success	790000.0
261	English for Career Development	760000.0
765	Successful Negotiation: Essential Strategies a...	750000.0
199	Data Science: Foundations using R	740000.0
211	Deep Learning	690000.0

Output 11

```
ct_enrolled=df.loc[:,['course_title', 'course_students_enrolled']]
ct_enrolled=ct_enrolled.sort_values(by=['course_students_enrolled'], ascending=False)
ct_enrolled_top10=ct_enrolled.head(10)
ct_enrolled_top10
```

Code 3

Let's analyse the titles on the basis of different sections like Data_Science, Law, AI, Statistics etc. Here, [*Output 12*] indicates the bar graph showing number of courses Titles from the particular subsection, the list of subsection includes *Causality, Data_Science, Law, AI, AWS, Statistics, Blockchain, Big_Data, Programming, Python, Data_Analysis, Data_Visualization, Health, English, Excel, Finance, Web_Development, Cloud Computing, Psychology and Cyber Security.* And by looking at the bar graph we can conclude that there are nearly 50 courses available on the Artificial Intelligence, followed by Health[38] & Python[35]. The code required to get the below figure is given as [*code 4*].

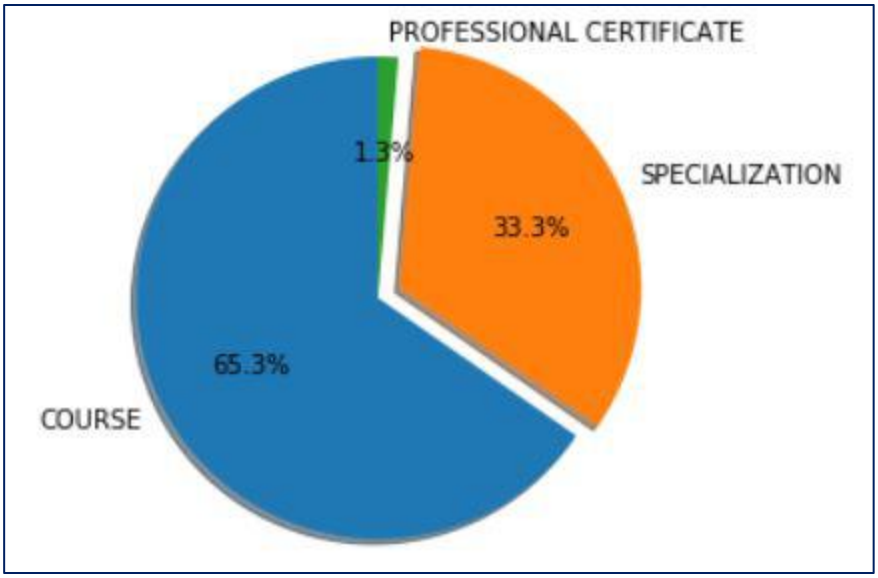


Output 12

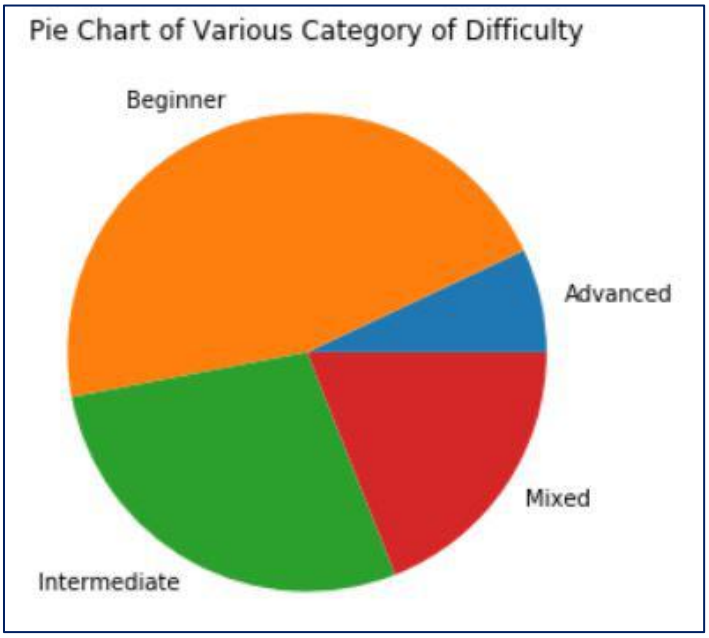
```
Causality=df["course_title"].str.contains("Causality",case=False).value_counts(ascending=True).to_frame()
Data_Science=df["course_title"].str.contains("Data Science",case=False).value_counts(ascending=True).to_frame()
Law=df["course_title"].str.contains("Law",case=False).value_counts(ascending=True).to_frame()
AI=df["course_title"].str.contains("AI",case=False).value_counts(ascending=True).to_frame()
AWS=df["course_title"].str.contains("AWS",case=False).value_counts(ascending=True).to_frame()
Statistics=df["course_title"].str.contains("Statistics",case=False).value_counts(ascending=True).to_frame()
Blockchain=df["course_title"].str.contains("Blockchain",case=False).value_counts(ascending=True).to_frame()
Big_Data=df["course_title"].str.contains("Big Data",case=False).value_counts(ascending=True).to_frame()
Programming=df["course_title"].str.contains("Programming",case=False).value_counts(ascending=True).to_frame()
python=df["course_title"].str.contains("python",case=False).value_counts(ascending=True).to_frame()
Data_Analysis=df["course_title"].str.contains("Data Analysis",case=False).value_counts(ascending=True).to_frame()
Data_Visualization=df["course_title"].str.contains("Data Visualization",case=False).value_counts(ascending=True).to_frame()
Health=df["course_title"].str.contains("Health",case=False).value_counts(ascending=True).to_frame()
English=df["course_title"].str.contains("English",case=False).value_counts(ascending=True).to_frame()
Excel=df["course_title"].str.contains("Excel",case=False).value_counts(ascending=True).to_frame()
Finance=df["course_title"].str.contains("Finance",case=False).value_counts(ascending=True).to_frame()
Web_Development=df["course_title"].str.contains("Web Development",case=False).value_counts(ascending=True).to_frame()
Cloud_Computing=df["course_title"].str.contains("Cloud Computing",case=False).value_counts(ascending=True).to_frame()
Psychology=df["course_title"].str.contains("Psychology",case=False).value_counts(ascending=True).to_frame()
Cyber_Security=df["course_title"].str.contains("Cyber Security",case=False).value_counts(ascending=True).to_frame()

result = pd.concat([Causality,Data_Science,Law,AI,AWS,Statistics,Blockchain,Big_Data,Programming,python,Data_Analysis,
                                    Data_Visualization,Health,English,Excel,Finance,Web_Development,Cloud_Computing,
                                    Psychology,Cyber_Security],axis=1).drop(index=0)
result_t=result.transpose()
result_t.columns=['count']
result_t.insert(0, "Names",["Causality","Data_Science","Law","AI","AWS","Statistics","Blockchain","Big_Data","Programming",
                                    "python","Data_Analysis","Data_Visualization","Health","English","Excel","Finance",
                                    "Web_Development","Cloud_Computing","Psychology","Cyber_Security"], True)
f,ax=plt.subplots(1,1,figsize=(25,7))
ax = sns.barplot(x="Names", y = "count", data = result_t)
ax.set_xticklabels(ax.get_xticklabels(), rotation=45,horizontalalignment='right',size=15)
```

Code 4



Output 13 : Course Certification Type[Pie Chart]



Output 14 : Course Difficulty Level[Pie Chart]

7. Conclusion

After doing this much analysis on the Coursera Dataset, I have concluded some points like,

I. It is very good platform for learning basics of anything as they offer 55% courses which are Beginner level.
II. For the Advanced Level Subjects/Courses you have to search for some other educational sites as coursera offers only 8% topic from advance level.
III. University of Pennesylvania is the leading Institute in terms of number of courses offered.
IV. Machine Learning is the Course enrolled by highest number of students.
V. AI is leading topic/subject with highest number of courses offered in that area.