PROBABLISTIC MODELING &REASONING WITH PYTHON PROJECT

```
PRESENTED BY:   FAUZIYA KHATOON
```

#TOPIC : QR WORLD RAKING UNIVERSITY

In [1]:

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

In [2]:

```
df=pd.read_csv("cwurData.csv")
```

In [3]:

```
df
```

Out[3]:

	world_rank	institution	country	national_rank	quality_of_education	alumni_emplo
0	1	Harvard University	USA	1	7	
1	2	Massachusetts Institute of Technology	USA	2	9	
2	3	Stanford University	USA	3	17	
3	4	University of Cambridge	United Kingdom	1	10	
4	5	California Institute of Technology	USA	4	2	
...	...	...	...	...	...	
2195	996	University of the Algarve	Portugal	7	367	
2196	997	Alexandria University	Egypt	4	236	
2197	998	Federal University of Ceará	Brazil	18	367	
2198	999	University of A Coruña	Spain	40	367	
2199	1000	China Pharmaceutical University	China	83	367	

2200 rows × 14 columns

# Taking Sampling

In [4]:

```
p= df.sample(620)
p
```

Out[4]:

	world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_
1911	712	San Francisco State University	USA	197	271	567	
1176	977	University of Puerto Rico at Mayagüez	Puerto Rico	1	355	478	
732	533	University of Udine	Italy	26	355	478	
2078	879	University of Orléans	France	43	367	524	
360	161	Goethe University Frankfurt	Germany	9	105	291	

# Exploratory Data Analysis (EDA)

In [5]:

```
p.head()                 # display first few rows of data frame
```

Out[5]:

	world_rank	institution	country	national_rank	quality_of_education	alumni_employme
1911	712	San Francisco State University	USA	197	271	56
1176	977	University of Puerto Rico at Mayagüez	Puerto Rico	1	355	47
732	533	University of Udine	Italy	26	355	47
2078	879	University of Orléans	France	43	367	52
360	161	Goethe University Frankfurt	Germany	9	105	29

In [6]:

```
p.tail()          # display the last few rows of a DataFrame.
```

Out[6]:

	world_rank	institution	country	national_rank	quality_of_education	alumni_employ
217	18	Swiss Federal Institute of Technology in Zurich	Switzerland	1	16	
1300	101	Technical University of Munich	Germany	3	37	
194	95	Tohoku University	Japan	6	43	
1768	569	King Saud University	Saudi Arabia	1	367	
1249	50	Rutgers University-New Brunswick	USA	33	91	

In [7]:

```
p.describe()      #  to display statistics discription  of numeric columns of a DataFrame
```

Out[7]:

	world_rank	national_rank	quality_of_education	alumni_employment	quality_of_faculty
count	620.000000	620.000000	620.000000	620.000000	620.000000
mean	462.622581	44.172581	273.316129	349.674194	178.482258
std	298.855982	55.967861	121.333530	183.381371	65.283609
min	1.000000	1.000000	1.000000	1.000000	1.000000
25%	195.250000	6.000000	178.750000	177.750000	176.500000
50%	470.500000	22.500000	355.000000	418.500000	210.000000
75%	721.500000	51.250000	367.000000	478.000000	218.000000
max	995.000000	225.000000	367.000000	567.000000	218.000000

In [8]:

```
p.info()      #  to display information about column data types and missing values.
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 620 entries, 1911 to 1249
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   world_rank           620 non-null    int64  
 1   institution          620 non-null    object 
 2   country              620 non-null    object 
 3   national_rank        620 non-null    int64  
 4   quality_of_education 620 non-null    int64  
 5   alumni_employment    620 non-null    int64  
 6   quality_of_faculty   620 non-null    int64  
 7   publications         620 non-null    int64  
 8   influence            620 non-null    int64  
 9   citations            620 non-null    int64  
 10  broad_impact         569 non-null    float64
 11  patents              620 non-null    int64  
 12  score                620 non-null    float64
 13  year                 620 non-null    int64  
dtypes: float64(2), int64(10), object(2)
```

SORTING

In [9]:

```
p.sort_index()      #  this method is used to sort the rows of a DataFrame
```

Out[9]:

	world_rank	institution	country	national_rank	quality_of_education	alumni_employm
2	3	Stanford University	USA	3	17	
7	8	Yale University	USA	6	14	
10	11	University of Chicago	USA	9	15	
12	13	University of Pennsylvania	USA	11	31	
16	17	Kyoto University	Japan	2	42	
...	...	...	...	...	...	
2182	983	Feng Chia University	Taiwan	21	367	
2184	985	Novosibirsk State University	Russia	5	167	
2188	989	University of Pau and Pays de l'Adour	France	49	367	
2190	991	Xidian University	China	81	367	
2194	995	King Abdulaziz University	Saudi Arabia	4	367	

620 rows × 14 columns

DATA CLEANING

In [10]:

```
p.dropna()        # used to handle missing values in data frame
```

Out[10]:

	world_rank	institution	country	national_rank	quality_of_education	alumni_employı
1911	712	San Francisco State University	USA	197	271	
1176	977	University of Puerto Rico at Mayagüez	Puerto Rico	1	355	
732	533	University of Udine	Italy	26	355	
2078	879	University of Orléans	France	43	367	
360	161	Goethe University Frankfurt	Germany	9	105	
...	...	...	...	...	...	
2056	857	University of Siegen	Germany	53	367	
217	18	Swiss Federal Institute of Technology in Zurich	Switzerland	1	16	
1300	101	Technical University of Munich	Germany	3	37	
1768	569	King Saud University	Saudi Arabia	1	367	
1249	50	Rutgers University-New Brunswick	USA	33	91	

569 rows × 14 columns

In [11]:

```
p.drop_duplicates()      #  used to remove duplicate rows
```

Out[11]:

	world_rank	institution	country	national_rank	quality_of_education	alumni_employı
1911	712	San Francisco State University	USA	197	271	
1176	977	University of Puerto Rico at Mayagüez	Puerto Rico	1	355	
732	533	University of Udine	Italy	26	355	
2078	879	University of Orléans	France	43	367	
360	161	Goethe University Frankfurt	Germany	9	105	
...	...	...	...	...	...	
217	18	Swiss Federal Institute of Technology in Zurich	Switzerland	1	16	
1300	101	Technical University of Munich	Germany	3	37	
194	95	Tohoku University	Japan	6	43	
1768	569	King Saud University	Saudi Arabia	1	367	
1249	50	Rutgers University-New Brunswick	USA	33	91	

620 rows × 14 columns

DATA VISUALIZATION

In [12]:

```
d=pd.DataFrame(p['year'].value_counts())     # calculate the frequency of occurance of ur
d.reset_index(inplace=True)                   # resets  the index of dataframe
d
```

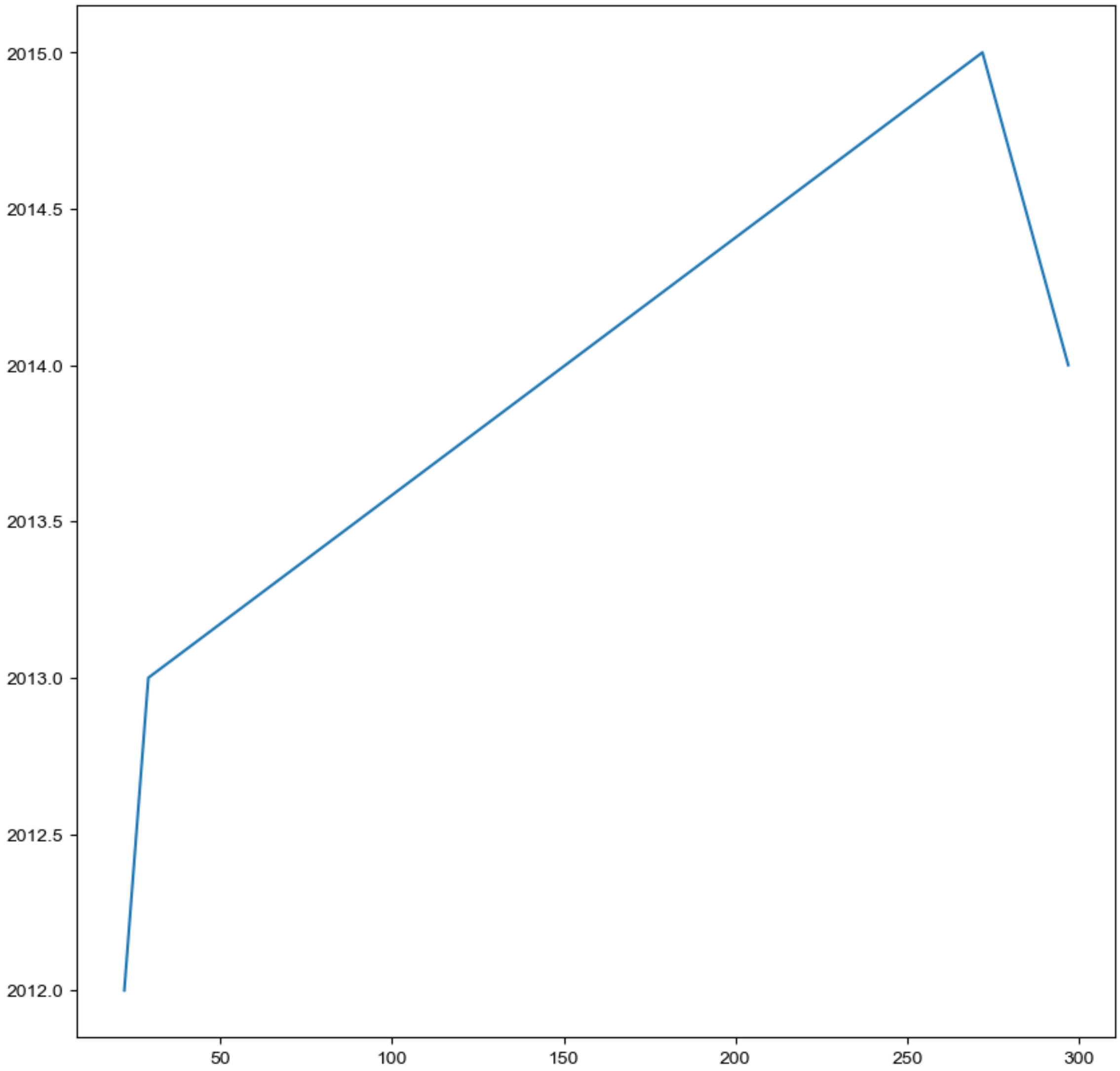
Out[12]:

	index	year
0	2014	297
1	2015	272
2	2013	29
3	2012	22

In [13]:

```
plt.figure(figsize=(10,10,))
plt.plot(d["year"],d["index"])
plt.style.use("seaborn-white")
```

```
C:\Users\fauzi\AppData\Local\Temp\ipykernel_2692\755744755.py:3: Matplotli
bDeprecationWarning: The seaborn styles shipped by Matplotlib are deprecat
ed since 3.6, as they no longer correspond to the styles shipped by seabor
n. However, they will remain available as 'seaborn-v0_8-<style>'. Alternat
ively, directly use the seaborn API instead.
  plt.style.use("seaborn-white")
```



In [14]:

```
d=pd.DataFrame(p['country'].value_counts())
d.reset_index(inplace=True)
d
```

Out[14]:

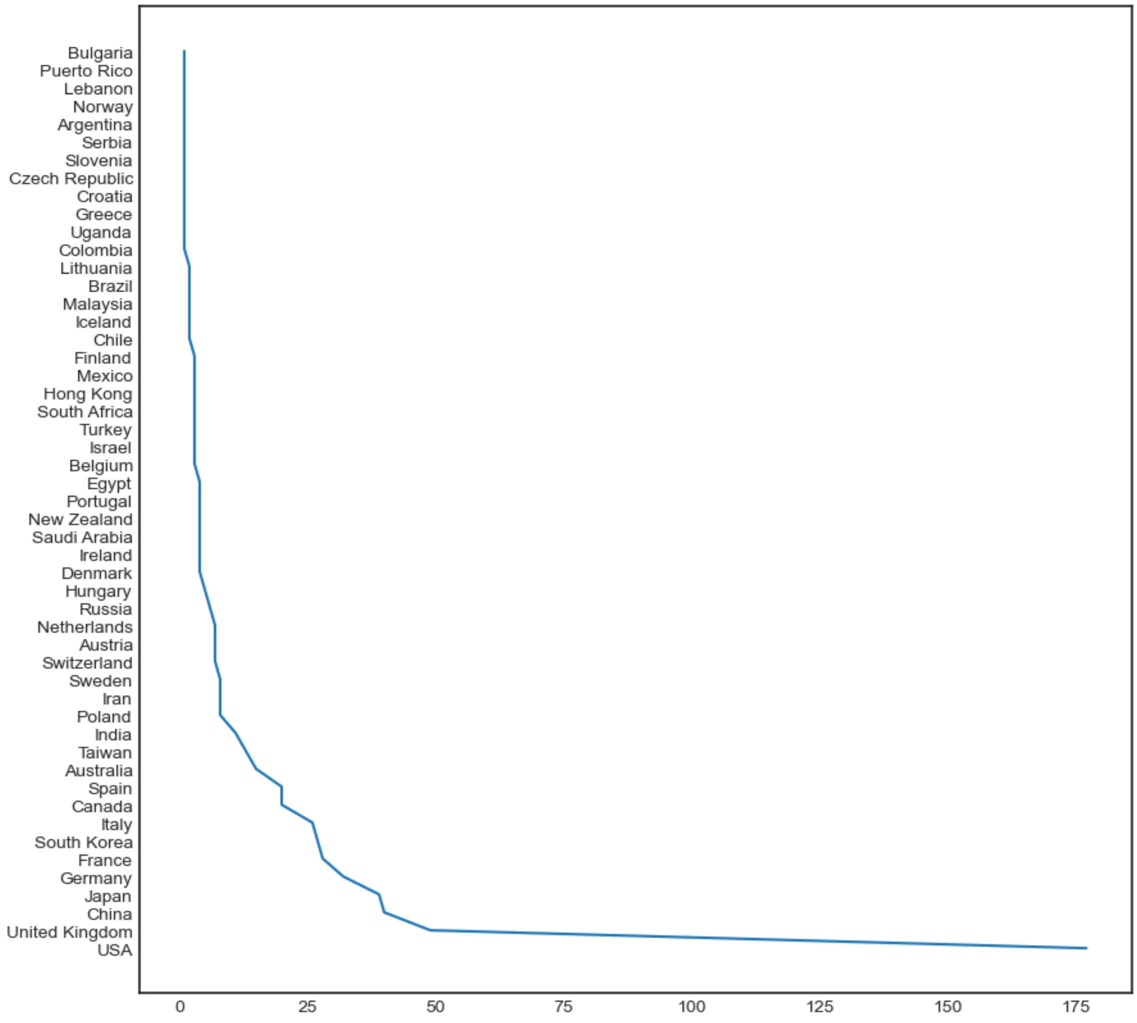
	index	country
0	USA	177
1	United Kingdom	49
2	China	40
3	Japan	39
4	Germany	32
5	France	28
6	South Korea	27
7	Italy	26
8	Canada	20
9	Spain	20
10	Australia	15
11	Taiwan	13
12	India	11
13	Poland	8
14	Iran	8
15	Sweden	8
16	Switzerland	7
17	Austria	7
18	Netherlands	7
19	Russia	6
20	Hungary	5
21	Denmark	4
22	Ireland	4
23	Saudi Arabia	4
24	New Zealand	4
25	Portugal	4
26	Egypt	4
27	Belgium	3
28	Israel	3
29	Turkey	3
30	South Africa	3
31	Hong Kong	3
32	Mexico	3
33	Finland	3
34	Chile	2
35	Iceland	2
36	Malaysia	2

	index	country
37	Brazil	2
38	Lithuania	2
39	Colombia	1
40	Uganda	1
41	Greece	1
42	Croatia	1
43	Czech Republic	1
44	Slovenia	1
45	Serbia	1
46	Argentina	1
47	Norway	1
48	Lebanon	1
49	Puerto Rico	1
50	Bulgaria	1

In [15]:

```
plt.figure(figsize=(10,10,))
plt.plot(d["country"],d["index"])
plt.style.use("seaborn-white")
```

```
C:\Users\fauzi\AppData\Local\Temp\ipykernel_2692\1786630107.py:3: Matplotl
ibDeprecationWarning: The seaborn styles shipped by Matplotlib are depreca
ted since 3.6, as they no longer correspond to the styles shipped by seabo
rn. However, they will remain available as 'seaborn-v0_8-<style>'. Alterna
tively, directly use the seaborn API instead.
  plt.style.use("seaborn-white")
```

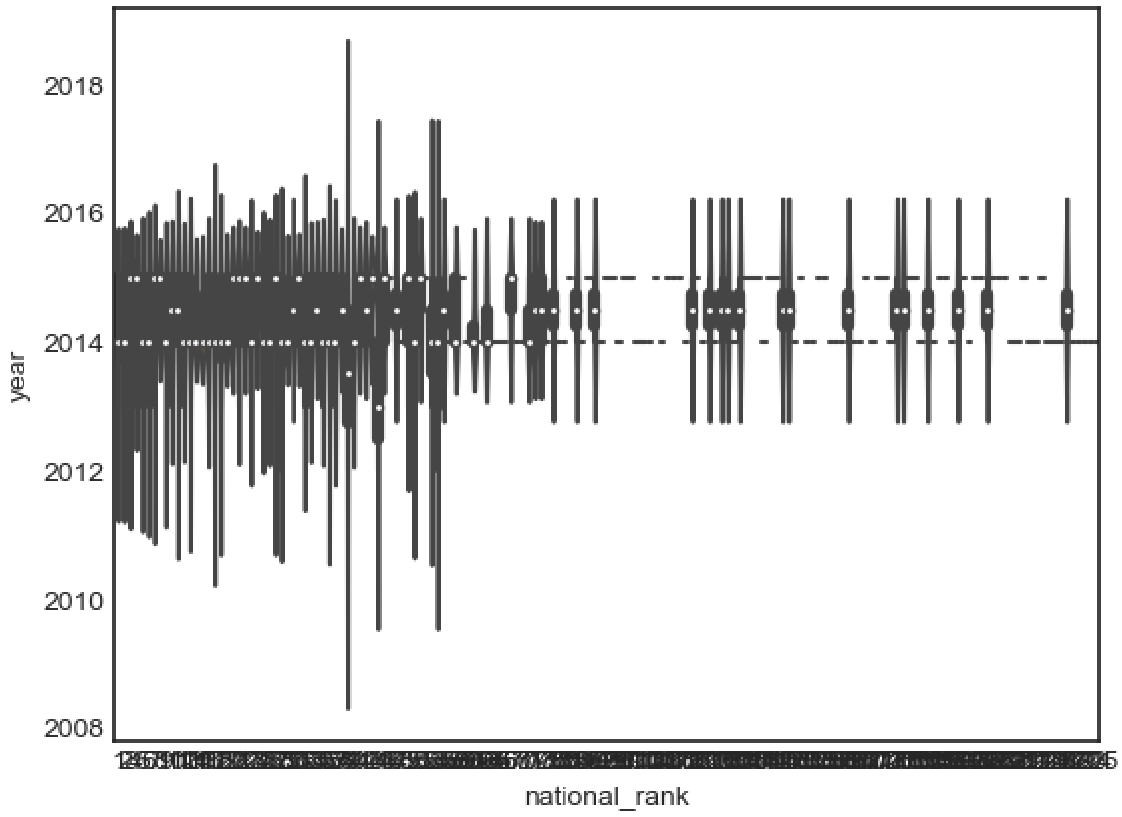


In [16]:

```
sns.violinplot(x='national_rank', y='year', data=p)
```

Out[16]:

```
<Axes: xlabel='national_rank', ylabel='year'>
```

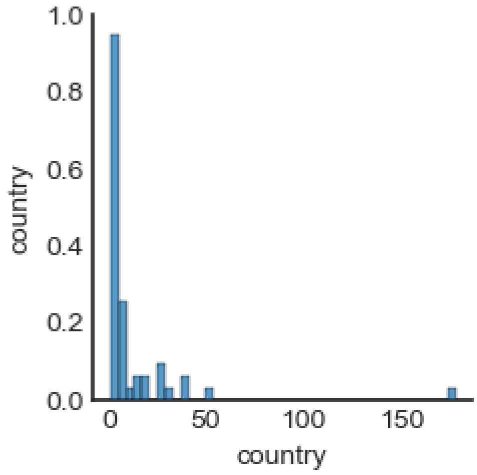


In [17]:

```
sns.pairplot(data=d)
```

Out[17]:

```
<seaborn.axisgrid.PairGrid at 0x21ff84c4af0>
```



In [18]:

```
a=p.sample(100)
a
```

Out[18]:

	world_rank	institution	country	national_rank	quality_of_education	alumni_employme
1033	834	Chung-Ang University	South Korea	30	355	47
772	573	Shizuoka University	Japan	35	355	47
721	522	SUNY Downstate Medical Center	USA	166	355	47
1823	624	Tokyo University of Agriculture and Technology	Japan	38	367	56
1350	151	University of Montreal	Canada	7	320	37
...	...	...	...	...	...	
702	503	Bar-Ilan University	Israel	6	266	40
1857	658	Massey University	New Zealand	5	367	49
754	555	Nagoya City University	Japan	33	355	30
2035	836	University of Regina	Canada	27	367	56
805	606	Binghamton University	USA	178	355	22

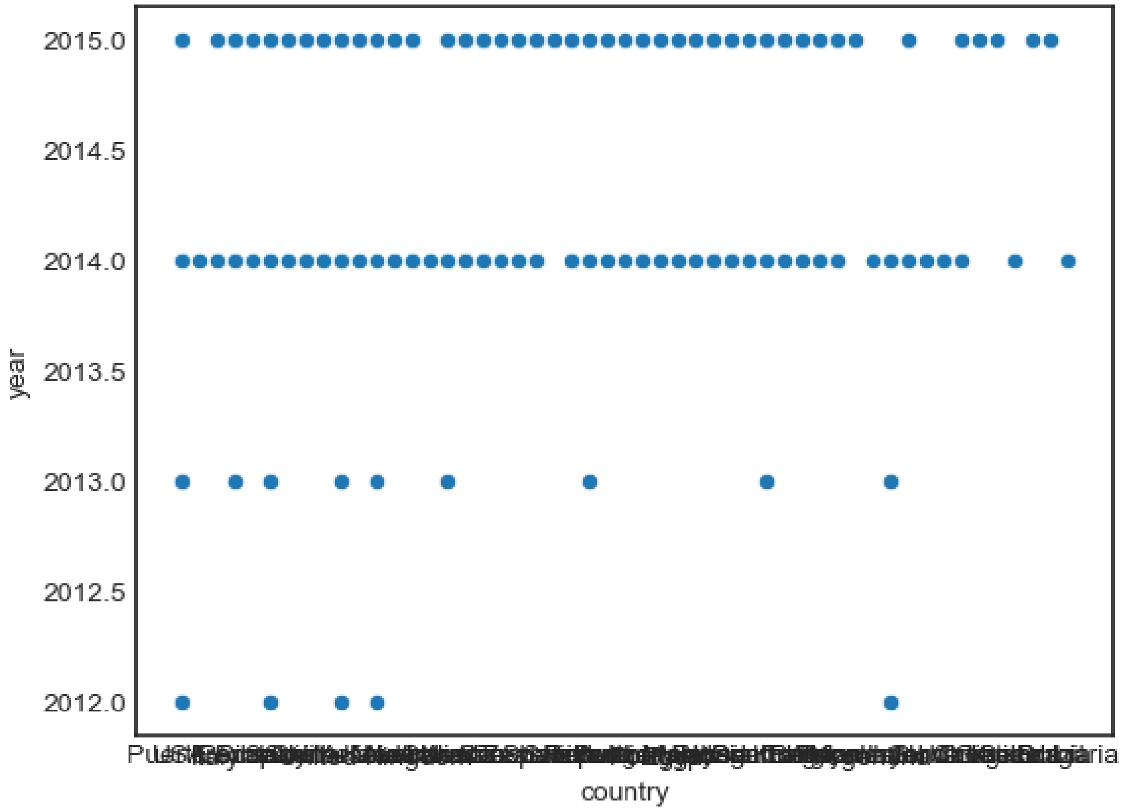
100 rows × 14 columns

In [19]:

```
sns.scatterplot(x='country', y='year', data=p)
```

Out[19]:

```
<Axes: xlabel='country', ylabel='year'>
```

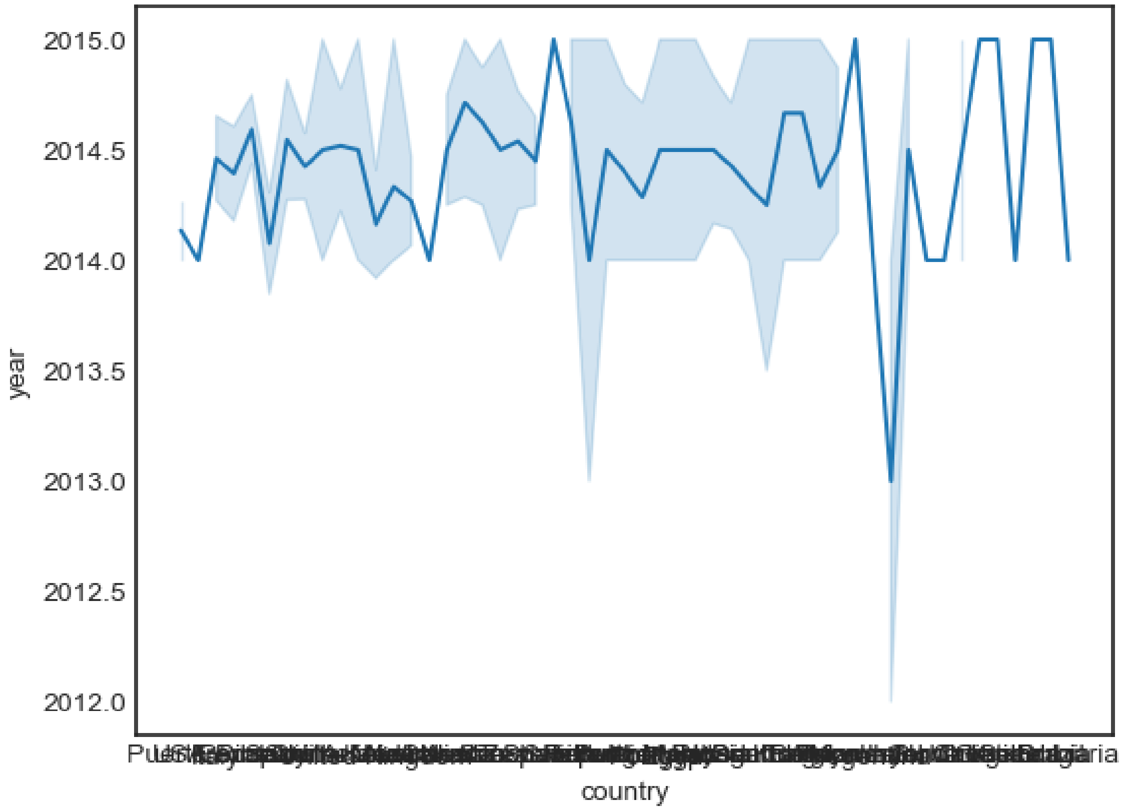


In [20]:

```
sns.lineplot(x='country', y='year', data=p)
```

Out[20]:

```
<Axes: xlabel='country', ylabel='year'>
```

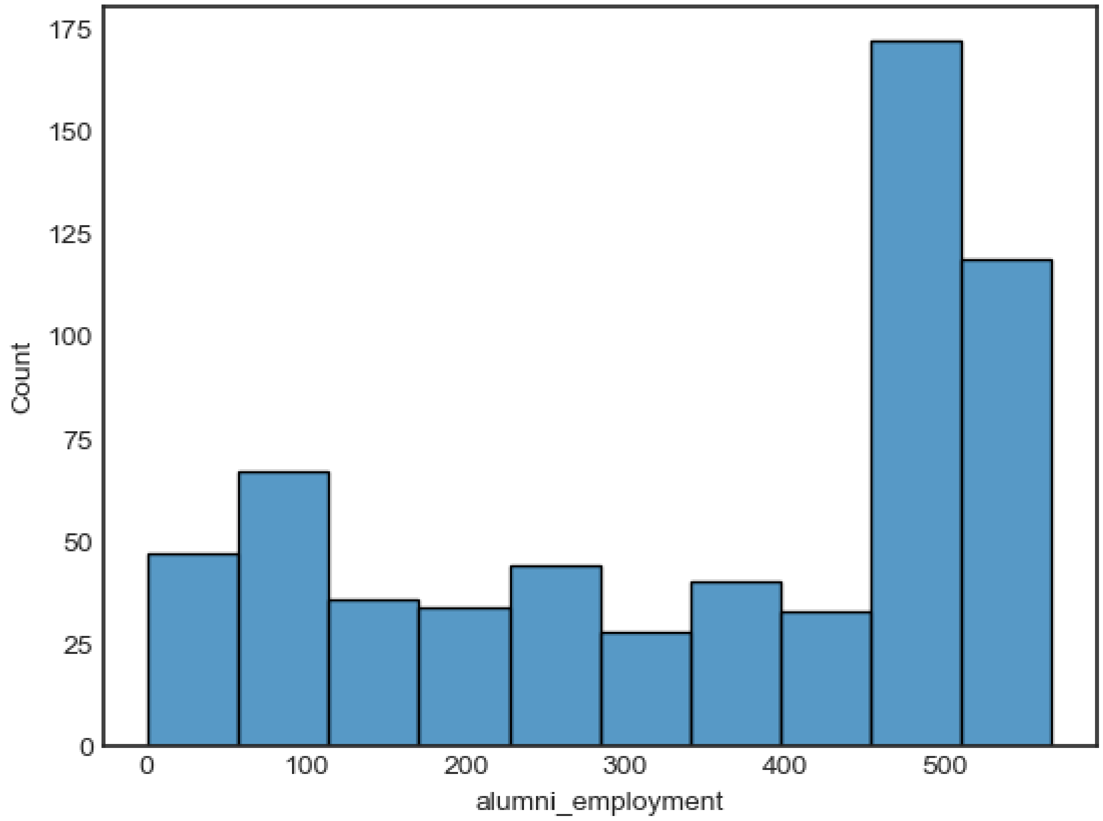


In [21]:

```
sns.histplot(data=p, x='alumni_employment', bins=10)
```

Out[21]:

```
<Axes: xlabel='alumni_employment', ylabel='Count'>
```

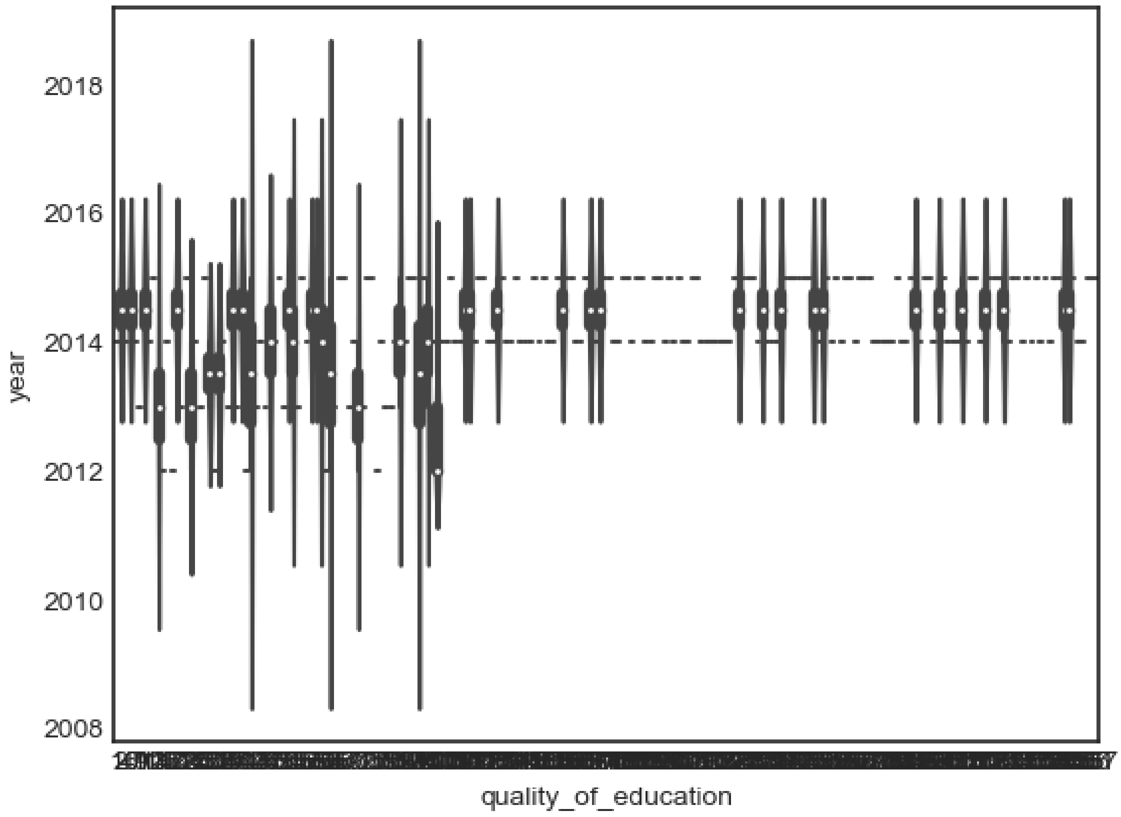


In [22]:

```
sns.violinplot(x='quality_of_education', y='year', data=p)
```

Out[22]:

```
<Axes: xlabel='quality_of_education', ylabel='year'>
```



THANK YOU !

PRESENTED BY : FAUZIYA KHATOON

In [ ]: