UDEMY COURSES ANALYSIS

Importing Libraries

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Importing Data

In [4]:
```
data = pd.read_csv('/content/udemy_courses.csv')
```

Checking datatypes

In [5]:
```
data.dtypes
```

Out[5]:
```
course_id               int64
course_title           object
url                    object
is_paid                  bool
price                   int64
num_subscribers         int64
num_reviews             int64
num_lectures            int64
level                  object
content_duration      float64
published_timestamp    object
subject                object
dtype: object
```

We need to change the datatype of a column published_timestamp from object to datetime

In [6]:
```
data['published_timestamp'] = pd.to_datetime(data['published_timestamp'])
```

In [7]:
```
data.dtypes
```

Out[7]:
```
course_id                            int64
course_title                        object
url                                 object
is_paid                               bool
price                                int64
num_subscribers                      int64
num_reviews                          int64
num_lectures                         int64
level                               object
content_duration                   float64
published_timestamp    datetime64[ns, UTC]
subject                             object
dtype: object
```

1. Display the Top 5 Rows of the Dataset

In [111…
```
data.head(5)
```

Out[111]:

	course_id	course_title	url	is_paid	price	num_subscribers	num_revie
0	1070968	Ultimate Investment Banking Course	https://www.udemy.com/ultimate-investment-bank...	True	200	2147	
1	1113822	Complete GST Course & Certification - Grow You...	https://www.udemy.com/goods-and-services-tax/	True	75	2792	
2	1006314	Financial Modeling for Business Analysts and C...	https://www.udemy.com/financial-modeling-for-b...	True	45	2174	
3	1210588	Beginner to Pro - Financial Analysis in Excel ...	https://www.udemy.com/complete-excel-finance-c...	True	95	2451	
4	1011058	How To Maximize Your Profits Trading Options	https://www.udemy.com/how-to-maximize-your-pro...	True	200	1276	

2. Check the Last 5 Rows of the Dataset

In [9]:
```
data.tail(5)
```

Out[9]:

	course_id	course_title	url	is_paid	price	num_subscribers	num_revie
3673	775618	Learn jQuery from Scratch - Master of JavaScri...	https://www.udemy.com/easy-jquery-for-beginner...	True	100	1040	
3674	1088178	How To Design A WordPress Website With No Codi...	https://www.udemy.com/how-to-make-a-wordpress-...	True	25	306	
3675	635248	Learn and Build using Polymer	https://www.udemy.com/learn-and-build-using-po...	True	40	513	
3676	905096	CSS Animations: Create Amazing Effects on Your...	https://www.udemy.com/css-animations-create-am...	True	50	300	
3677	297602	Using MODX CMS to Build Websites: A Beginner's...	https://www.udemy.com/using-modx-cms-to-build-...	True	45	901	

3. Find Shape of our dataset(Number of rows and Number of Columns)

In [10]:
```
data.shape
```

Out[10]:
```
(3678, 12)
```

In [11]:
```
print('Numer of Rows:',data.shape[0])
print('Number of Columns:',data.shape[1])
```

```
Numer of Rows: 3678
Number of Columns: 12
```

4. Getting information about dataset like the total number of rows, the total number of columns, datatypes of each column, and memory Requirements

In [12]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3678 entries, 0 to 3677
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   course_id            3678 non-null   int64
 1   course_title         3678 non-null   object
 2   url                  3678 non-null   object
 3   is_paid              3678 non-null   bool
 4   price                3678 non-null   int64
 5   num_subscribers      3678 non-null   int64
 6   num_reviews          3678 non-null   int64
 7   num_lectures         3678 non-null   int64
 8   level                3678 non-null   object
 9   content_duration     3678 non-null   float64
 10  published_timestamp  3678 non-null   datetime64[ns, UTC]
 11  subject              3678 non-null   object
dtypes: bool(1), datetime64[ns, UTC](1), float64(1), int64(5), object(4)
memory usage: 319.8+ KB
```

5. Check Null values in the dataset

In [13]:
```
data.isnull()
```

Out[13]:

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
3673	False	False	False	False	False	False	False	False	False
3674	False	False	False	False	False	False	False	False	False
3675	False	False	False	False	False	False	False	False	False
3676	False	False	False	False	False	False	False	False	False
3677	False	False	False	False	False	False	False	False	False

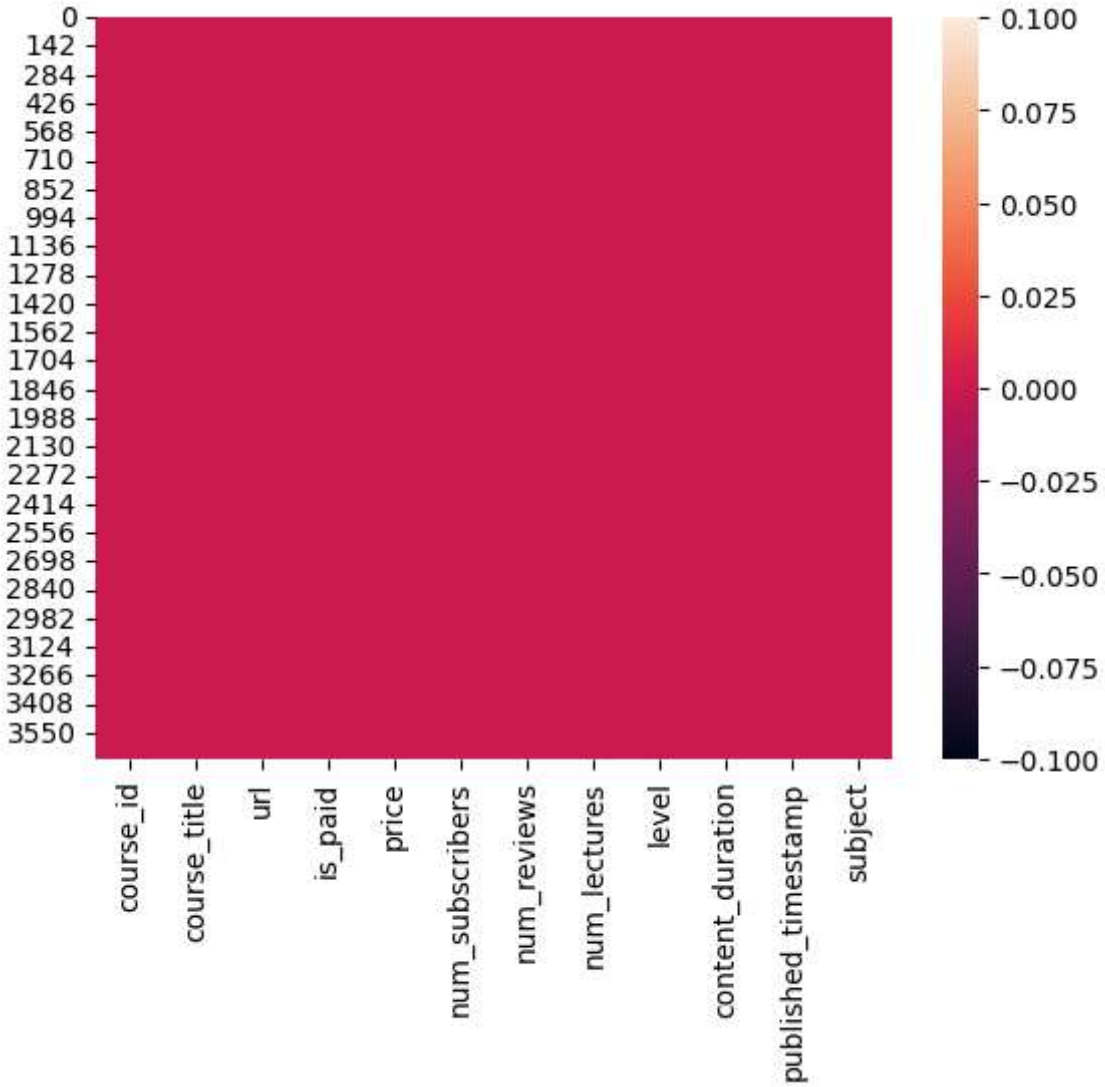
3678 rows × 12 columns

In [14]:
```
data.isnull().sum()
```

Out[14]:
```
course_id              0
course_title           0
url                    0
is_paid                0
price                  0
num_subscribers        0
num_reviews            0
num_lectures           0
level                  0
content_duration       0
published_timestamp    0
subject                0
dtype: int64
```

In [15]:
```
sns.heatmap(data.isnull())
```

Out[15]:
```
<Axes: >
```



6. Check for Duplicate data and drop them

In [16]:
```
dup = data.duplicated().any()
print('Any Duplicates:',dup)
```

```
Any Duplicates: True
```

In [17]:
```
data = data.drop_duplicates()
```

In [18]:
```
dup = data.duplicated().any()
print('Any Duplicates:',dup)
```

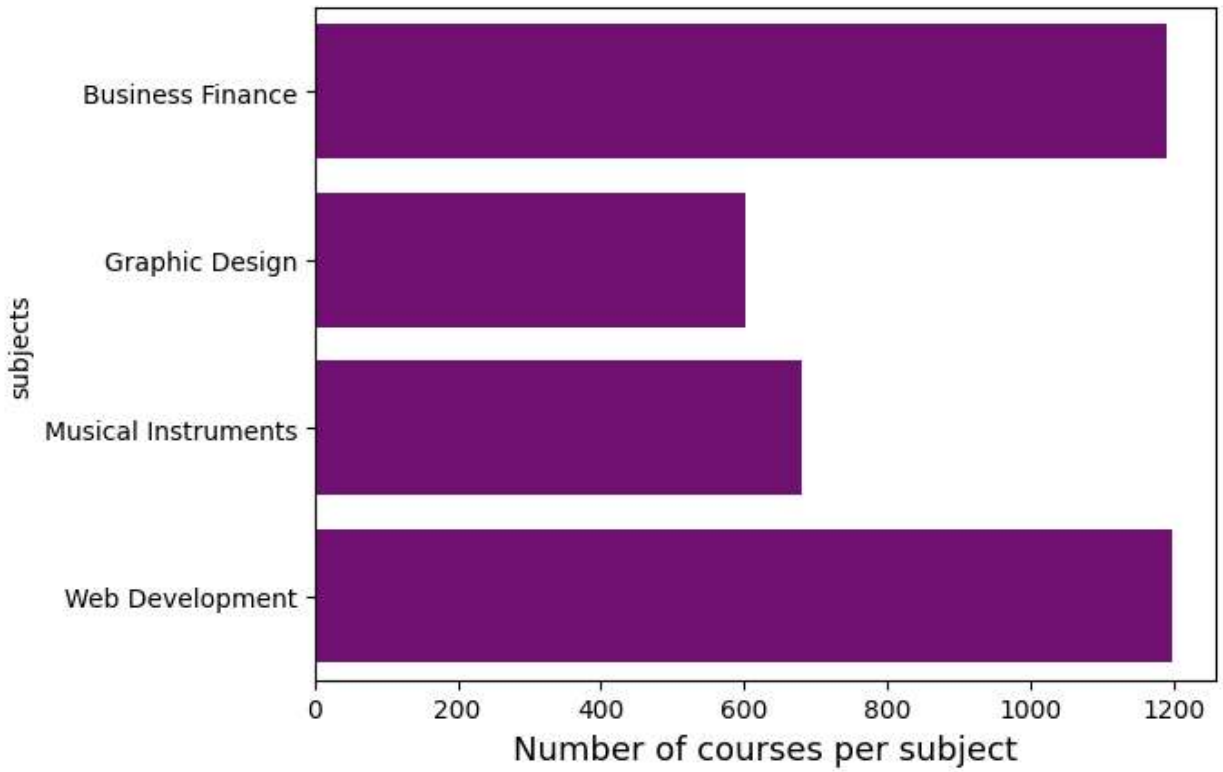
```
Any Duplicates: False
```

7. Find Out the number of courses per Subject

In [19]:
```
data['subject'].value_counts()
```

Out[19]:
```
Web Development        1199
Business Finance       1191
Musical Instruments     680
Graphic Design          602
Name: subject, dtype: int64
```

In [21]:
```
sns.countplot(data['subject'], color = 'purple')
plt.xlabel('Number of courses per subject',fontsize = 13)
plt.ylabel('subjects')
plt.show()
```

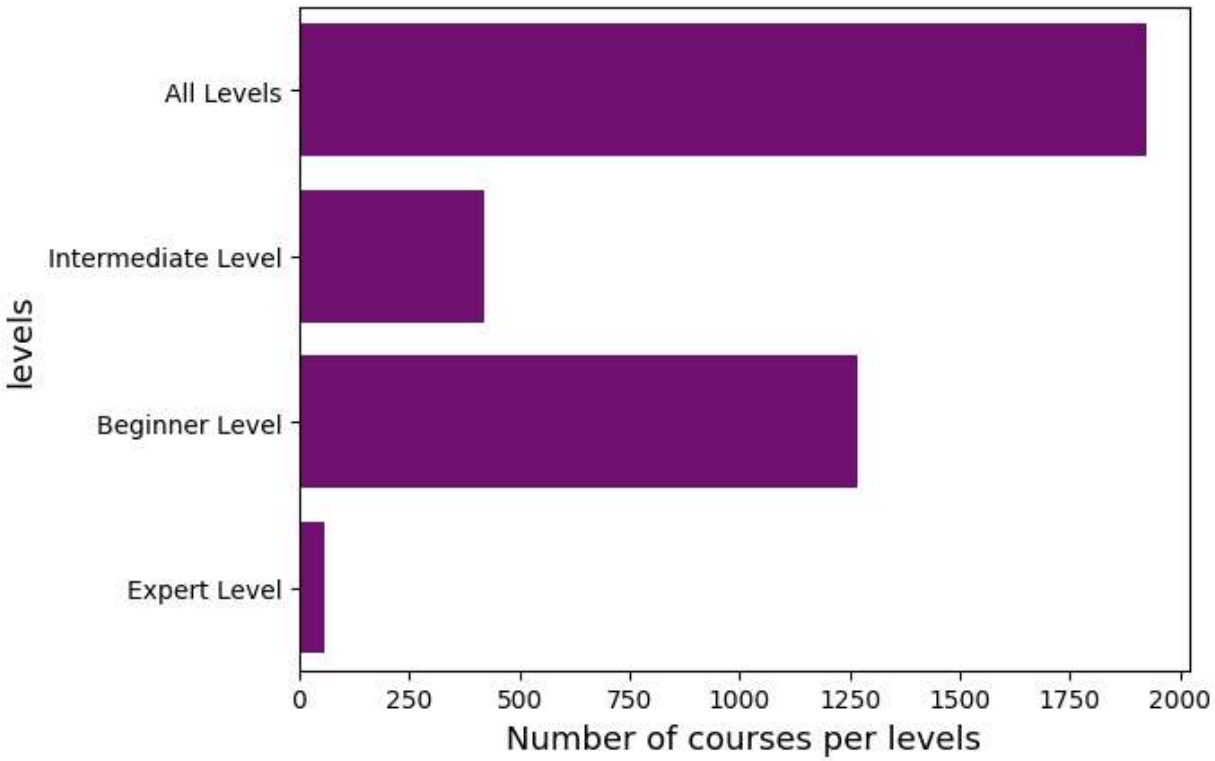


8. For which levels, Udemy courses providing the courses

In [22]:
```
data['level'].value_counts()
```

Out[22]:
```
All Levels            1925
Beginner Level        1268
Intermediate Level     421
Expert Level            58
Name: level, dtype: int64
```

In [23]:
```
sns.countplot(data['level'],color = 'purple')
plt.xlabel('Number of courses per levels',fontsize = 13)
plt.ylabel('levels',fontsize = 13)
plt.show()
```

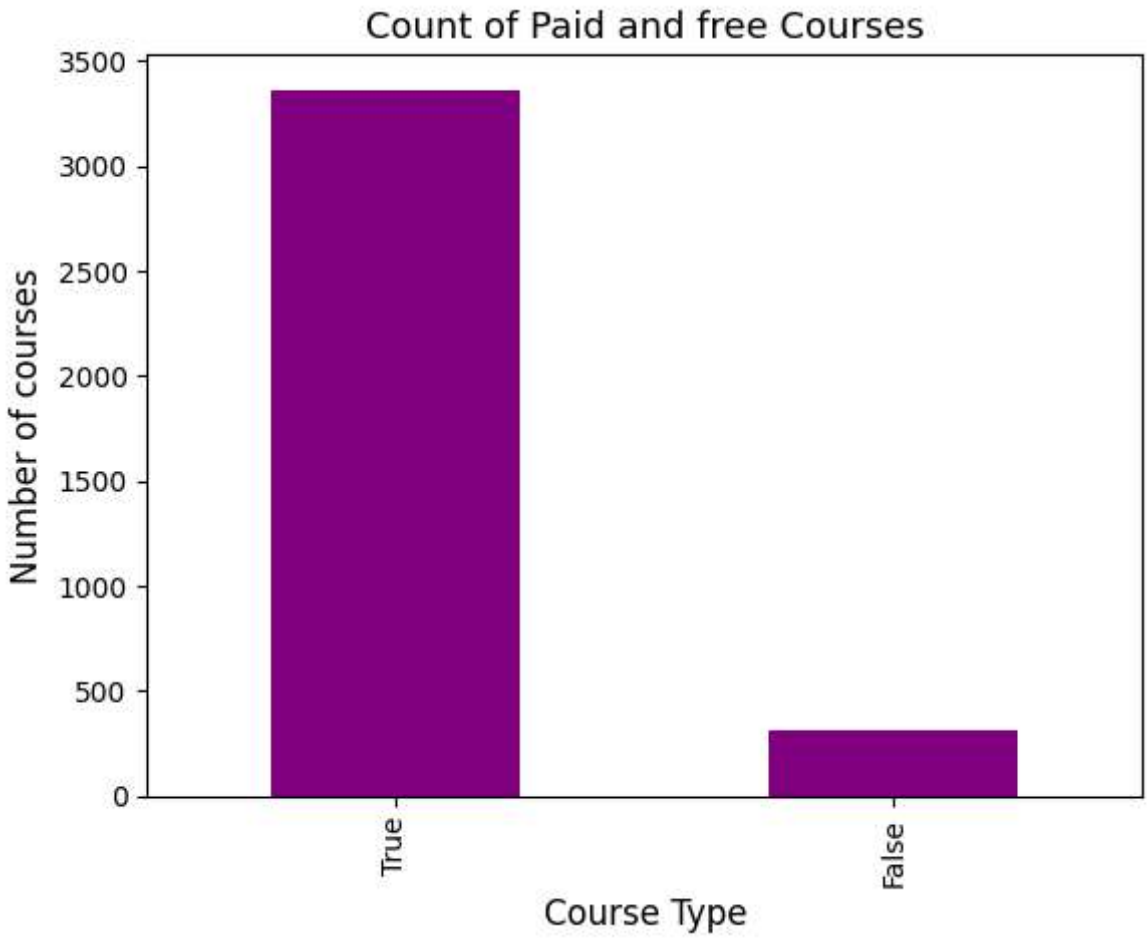


9. Display the Count of paid and free courses

In [24]:
```
data['is_paid'].value_counts()
```

Out[24]:
```
True     3362
False     310
Name: is_paid, dtype: int64
```

In [25]:
```
data['is_paid'].value_counts().plot(kind = 'bar',color = 'purple')
plt.title('Count of Paid and free Courses',fontsize = 13)
plt.xlabel('Course Type',fontsize = 12)
plt.ylabel('Number of courses',fontsize = 12)
plt.show()
```

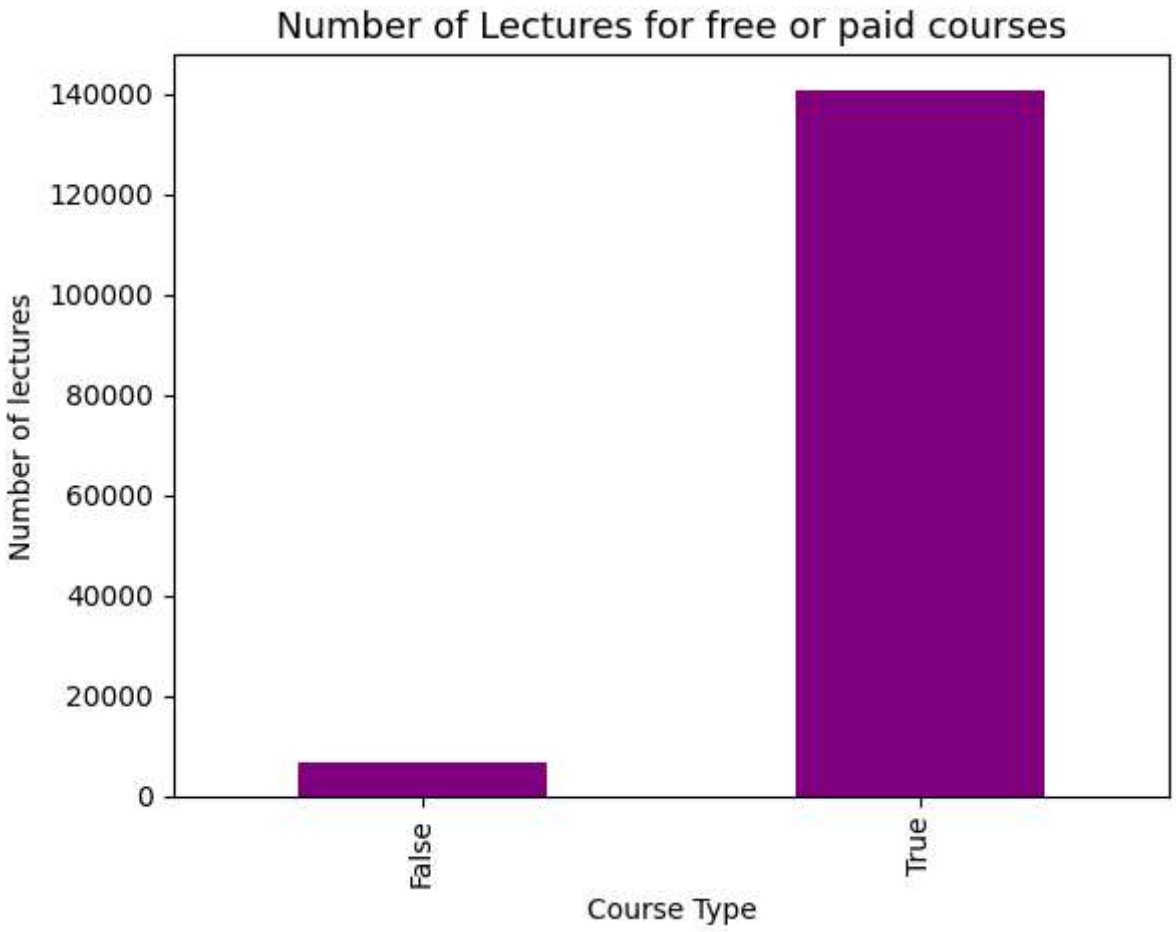


10. Which courses has more lectures(free or paid)?

In [26]:
```
lecture_count = data.groupby('is_paid')['num_lectures'].sum()
lecture_count
```

Out[26]:
```
is_paid
False      6639
True     140756
Name: num_lectures, dtype: int64
```

In [27]:
```
lecture_count.plot(kind = 'bar',color = 'purple')
plt.title('Number of Lectures for free or paid courses',fontsize = 13)
plt.xlabel('Course Type')
plt.ylabel('Number of lectures')
plt.show()
```



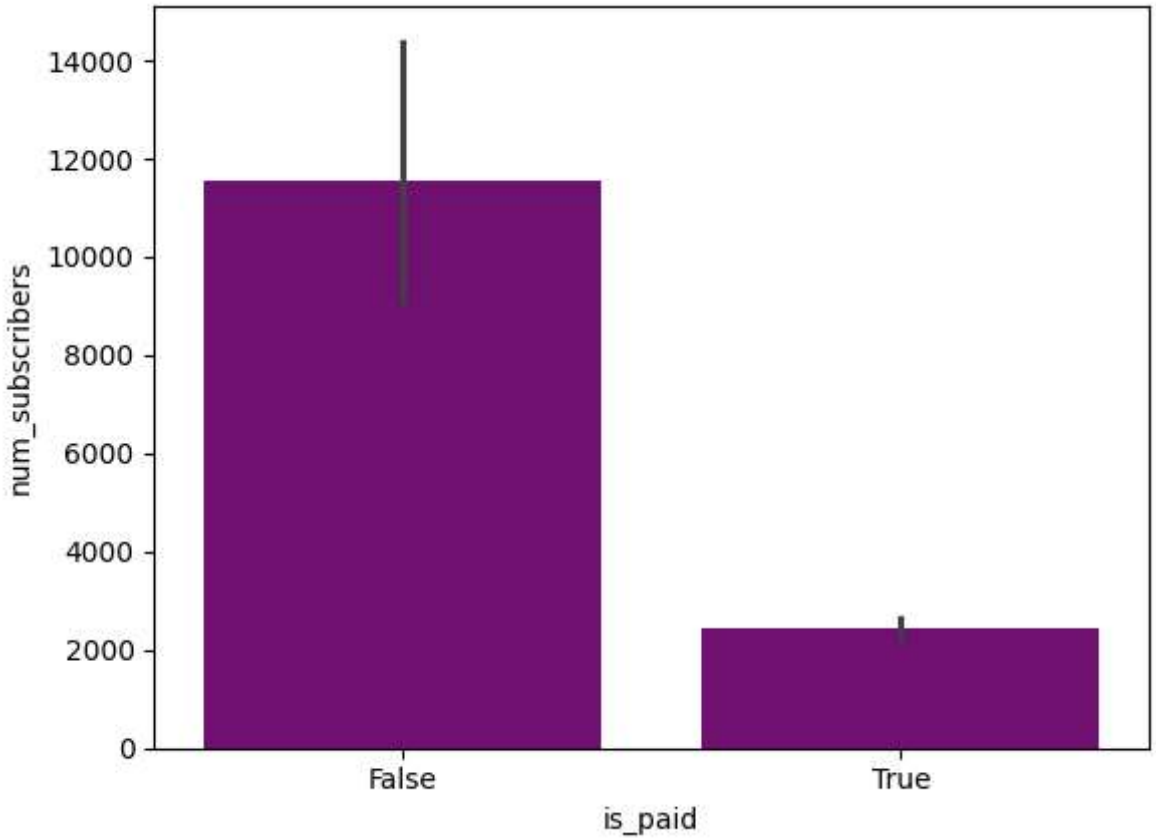
11. Which Courses Have A Higher Number of Subscribers Free or Paid?

In [29]:
```
list(data.columns)
```

Out[29]:
```
['course_id',
 'course_title',
 'url',
 'is_paid',
 'price',
 'num_subscribers',
 'num_reviews',
 'num_lectures',
 'level',
 'content_duration',
 'published_timestamp',
 'subject']
```

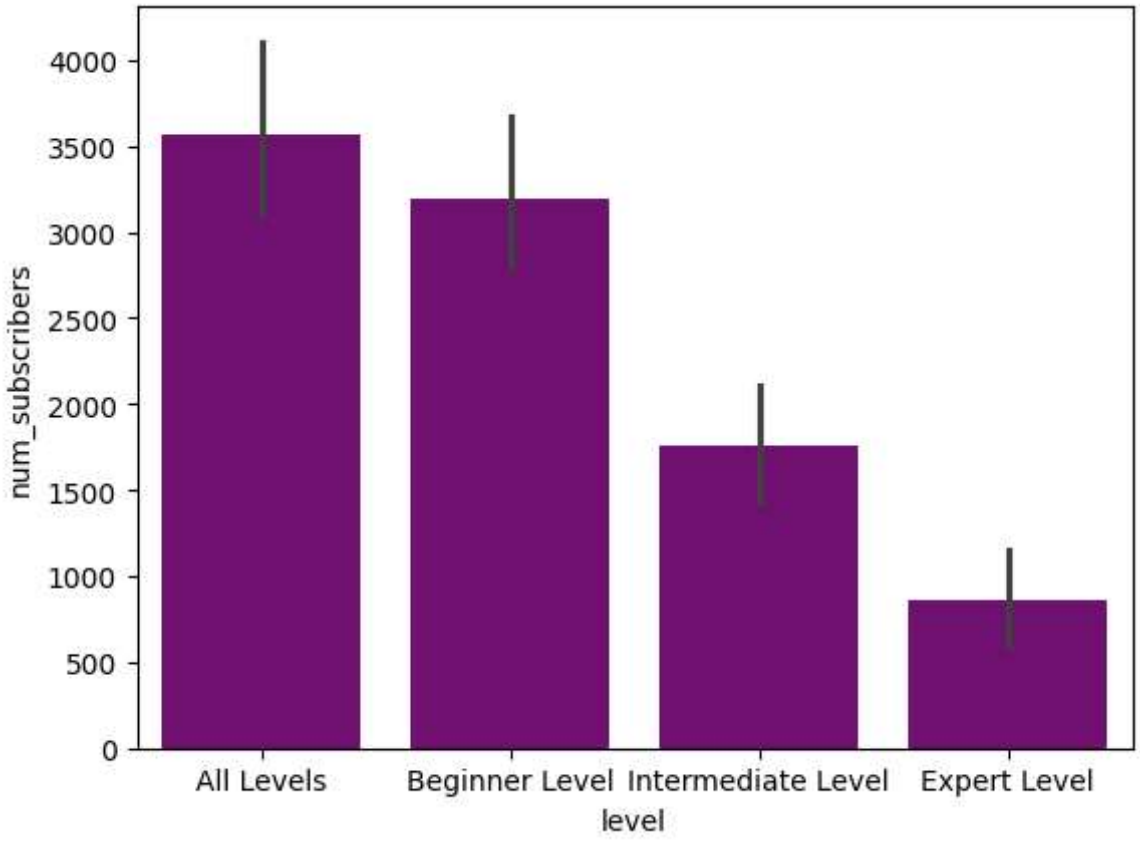
In [38]:
```
sns.barplot(x = 'is_paid',y = 'num_subscribers',data = data,color = 'purple')
```

Out[38]:
```
<Axes: xlabel='is_paid', ylabel='num_subscribers'>
```



12. Which Level Has The Highest Number of Subscribers?

In [43]:
```
sns.barplot(x = 'level', y = 'num_subscribers',data = data,color = 'purple',order = da
plt.show()
```



13. Find the Most Popular Course Title

In [45]:
```
data[data['num_subscribers'].max() == data['num_subscribers']]['course_title']
```

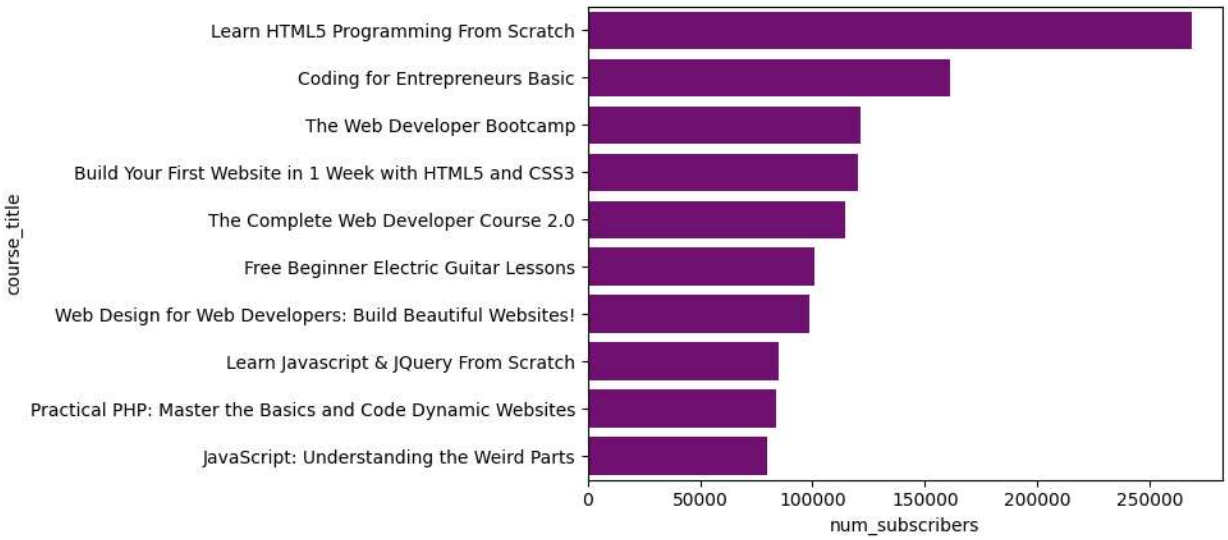
Out[45]:
```
2827    Learn HTML5 Programming From Scratch
Name: course_title, dtype: object
```

14. Display 10 Most Popular Courses As Per Number of Subscribers

In [ ]:
```
top_10 = data.sort_values(by = 'num_subscribers',ascending = False).head(10)
top_10
```

In [71]:
```
sns.barplot(x = 'num_subscribers',y = 'course_title',data = top_10,color = 'purple')
```

Out[71]:
```
<Axes: xlabel='num_subscribers', ylabel='course_title'>
```



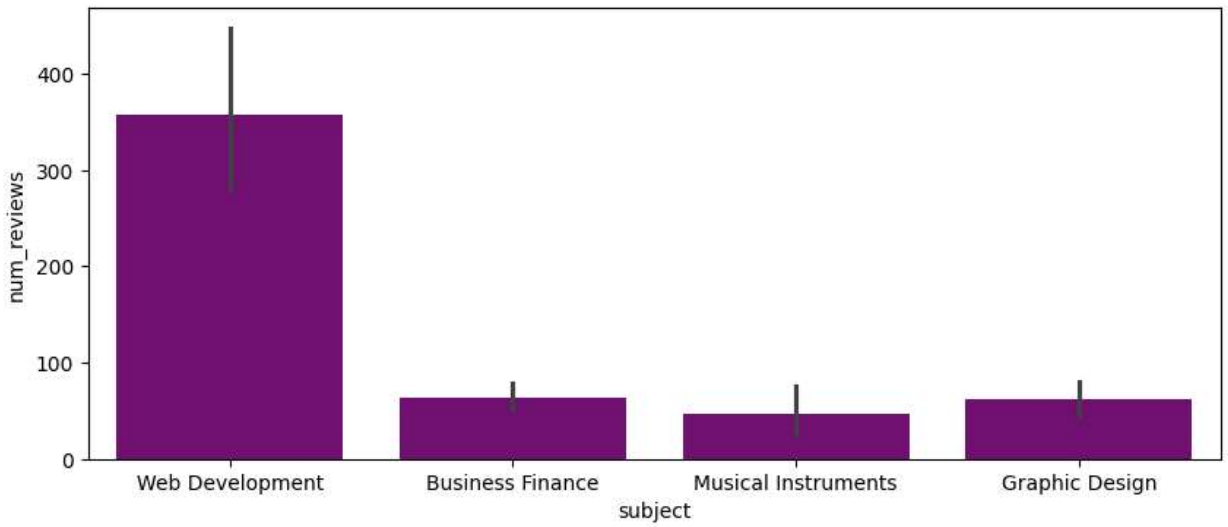
15. Find The Course Which Is Having The Highest Number of Reviews.

In [58]:
```
data[data['num_reviews'].max() == data['num_reviews']]['subject']
```

Out[58]:
```
3230    Web Development
Name: subject, dtype: object
```

In [68]:
```
plt.figure(figsize = (10,4))
sns.barplot(x = 'subject',y = 'num_reviews',data= data,color = 'purple',order = data['
```

Out[68]:
```
<Axes: xlabel='subject', ylabel='num_reviews'>
```



Top 5 courses with Highest Number of Reviews

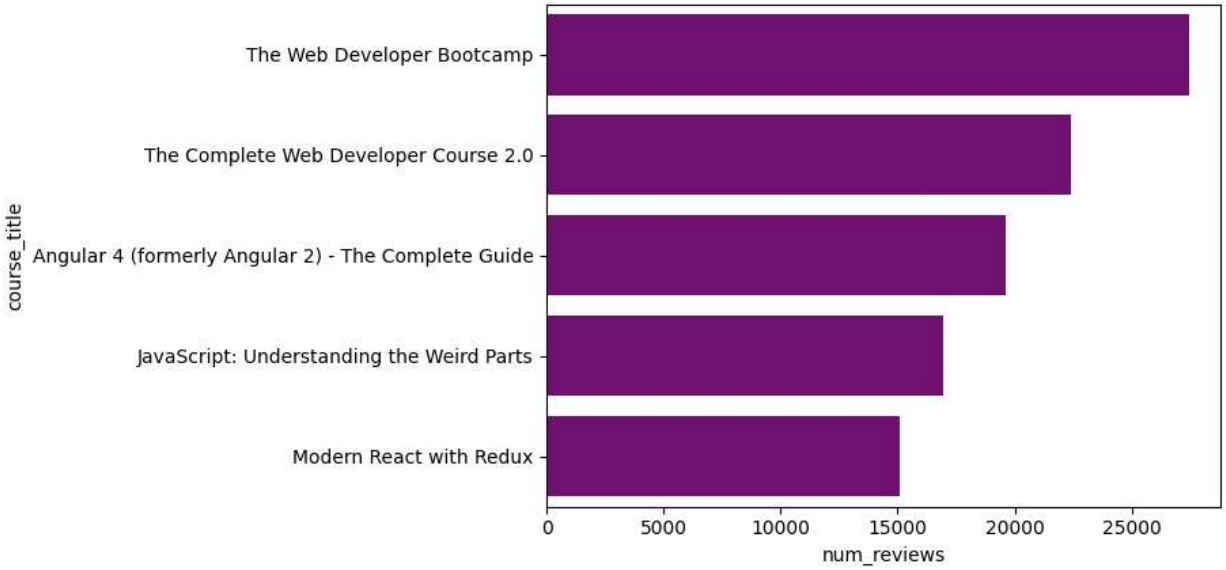
In [56]:
```
top_5 = data[['num_reviews', 'course_title']].sort_values(by = 'num_reviews',ascending
top_5
```

Out[56]:

	num_reviews	course_title
3230	27445	The Web Developer Bootcamp
3232	22412	The Complete Web Developer Course 2.0
3204	19649	Angular 4 (formerly Angular 2) - The Complete ...
3247	16976	JavaScript: Understanding the Weird Parts
3254	15117	Modern React with Redux

In [73]:
```
sns.barplot(x = 'num_reviews',y = 'course_title',data= top_5,color = 'purple')
```

Out[73]:
```
<Axes: xlabel='num_reviews', ylabel='course_title'>
```



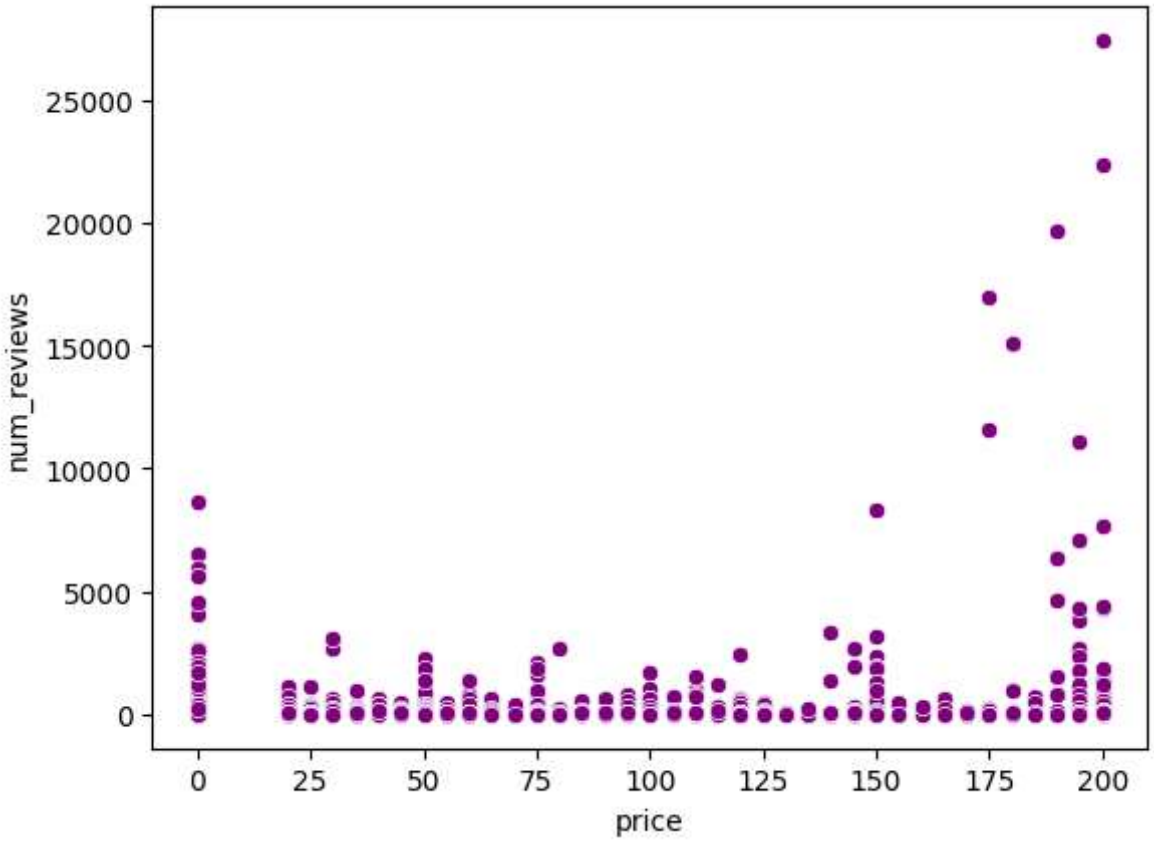
16. Does Price Affect the Number of Reviews?

In [74]:
```
list(data.columns)
```

Out[74]:
```
['course_id',
 'course_title',
 'url',
 'is_paid',
 'price',
 'num_subscribers',
 'num_reviews',
 'num_lectures',
 'level',
 'content_duration',
 'published_timestamp',
 'subject']
```

In [76]:
```
sns.scatterplot(x = 'price',y = 'num_reviews',data = data,color = 'purple')
```

Out[76]:
```
<Axes: xlabel='price', ylabel='num_reviews'>
```



17. Find the Total Number of Courses Related To Python

In [79]:
```
len(data[data['course_title'].str.contains('python',case = False)])
```

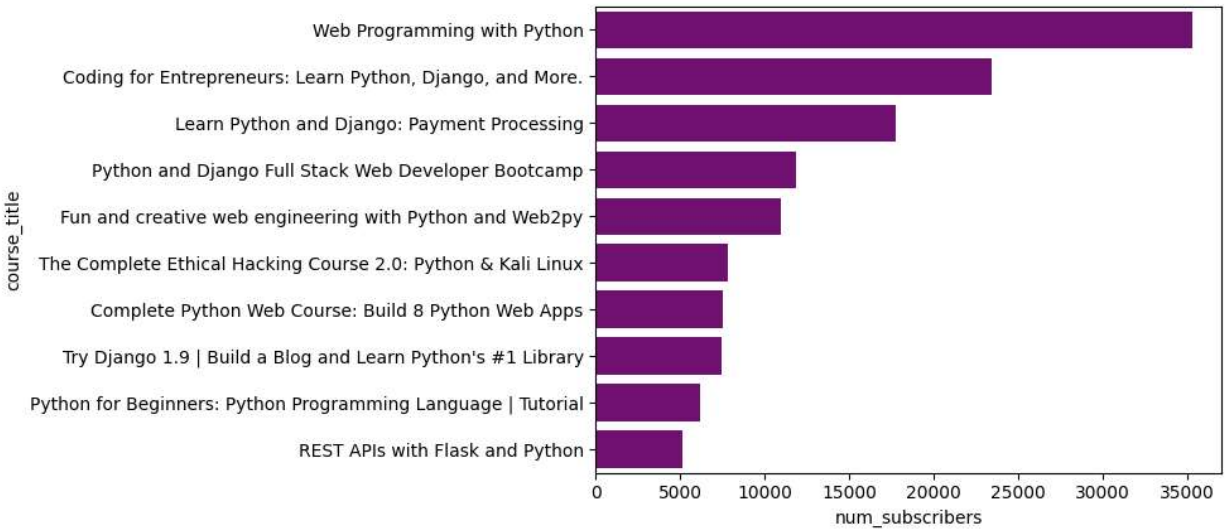
Out[79]:
```
29
```

18. Display 10 Most Popular Python Courses As Per Number of Subscribers

In [ ]:
```
python = data[data['course_title'].str.contains('python',case = False)]\
.sort_values(by = 'num_subscribers',ascending = False).head(10)
```

In [81]:
```
sns.barplot(x = 'num_subscribers',y = 'course_title',data = python,color = 'purple')
```

Out[81]:
```
<Axes: xlabel='num_subscribers', ylabel='course_title'>
```



19. In Which Year The Highest Number of Courses Were Posted?

In [83]:
```
list(data.columns)
```

Out[83]:
```
['course_id',
 'course_title',
 'url',
 'is_paid',
 'price',
 'num_subscribers',
 'num_reviews',
 'num_lectures',
 'level',
 'content_duration',
 'published_timestamp',
 'subject']
```

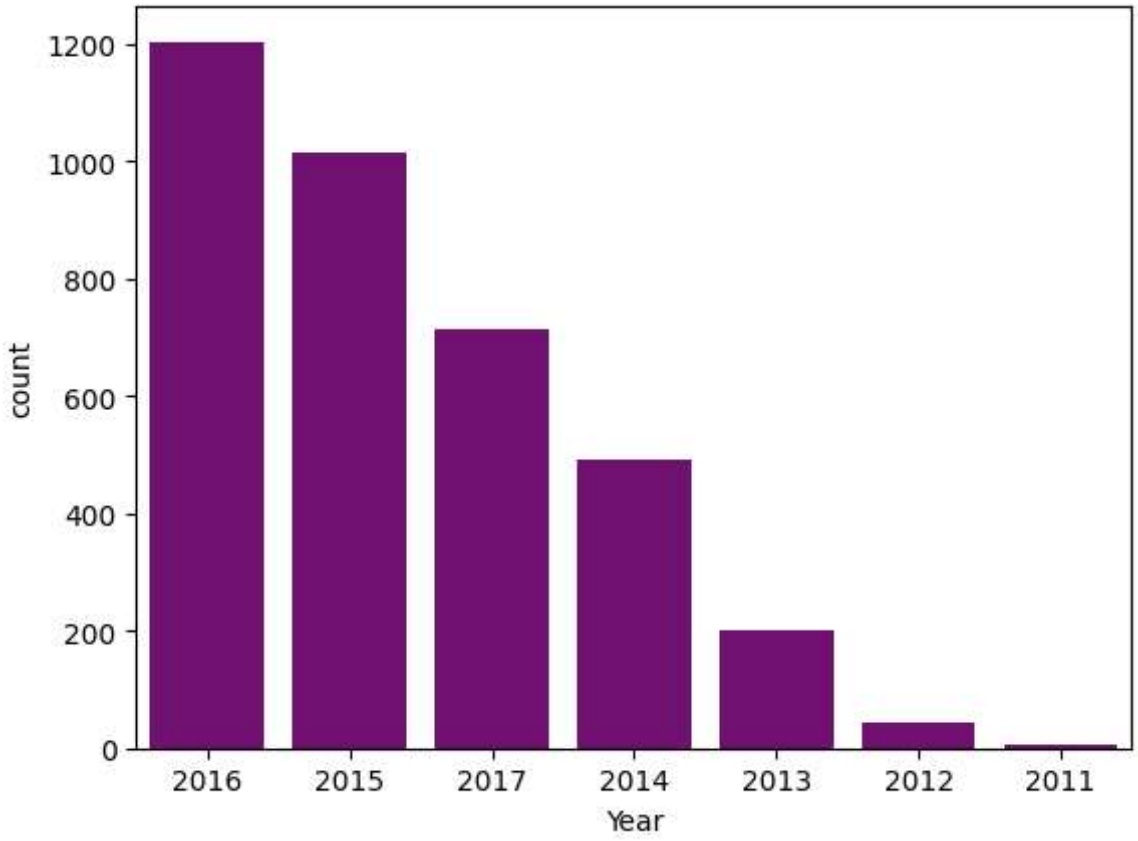
In [84]:
```
data['Year'] = data['published_timestamp'].dt.year
```

In [100…
```
Course_per_year = data.groupby('Year').size().sort_values(ascending = False)
Course_per_year
```

Out[100]:
```
Year
2016    1204
2015    1014
2017     713
2014     490
2013     201
2012      45
2011       5
dtype: int64
```

In [89]:
```
sns.countplot(x = 'Year',data = data,color = 'purple',order = data['Year'].value_count
```

Out[89]:
```
<Axes: xlabel='Year', ylabel='count'>
```



In [103…

```
Highest_Posted_course = Course_per_year.idxmax()
print('Highest Number of courses posted per year :',Highest_Posted_course)
```

```
Highest Number of courses posted per year : 2016
```

20. Display Category-Wise Count of Posted Subjects [Year Wise]

In [110…

```
data.groupby('Year')['subject'].value_counts()
```

Out[110]:

Year	subject	
2011	Web Development	5
2012	Web Development	19
	Graphic Design	10
	Musical Instruments	10
	Business Finance	6
2013	Business Finance	84
	Web Development	55
	Musical Instruments	39
	Graphic Design	23
2014	Business Finance	192
	Musical Instruments	120
	Web Development	113
	Graphic Design	65
2015	Business Finance	339
	Web Development	336
	Musical Instruments	171
	Graphic Design	168
2016	Web Development	448
	Business Finance	347
	Musical Instruments	228
	Graphic Design	181
2017	Business Finance	223
	Web Development	223
	Graphic Design	155
	Musical Instruments	112

Name: subject, dtype: int64