import pandas as pd import seaborn as sns import matplotlib.pyplot as plt

0. Dataset can be found on:

https://www.kaggle.com/datasets/andrewmvd/udemy-courses

In [4]:
```
# I would like to appreciate to Data Thinkers on Youtube and the google play store apps dataset owner
```

In [2]:
```
data = pd.read_csv(r'E:\Data Analyst Project\udemy_courses.csv',parse_dates=['published_timestamp'])
```

1. Show top rows of the Dataset

In [11]:
```
data.head(5)
```

Out[11]:

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_
0	1070968	Ultimate Investment Banking Course	https://www.udemy.com/ultimate-investment-bank...	True	200	2147	23	51	All Levels	
1	1113822	Complete GST Course & Certification - Grow You...	https://www.udemy.com/goods-and-services-tax/	True	75	2792	923	274	All Levels	
2	1006314	Financial Modeling for Business Analysts and C...	https://www.udemy.com/financial-modeling-for-b...	True	45	2174	74	51	Intermediate Level	
3	1210588	Beginner to Pro - Financial Analysis in Excel ...	https://www.udemy.com/complete-excel-finance-c...	True	95	2451	11	36	All Levels	
4	1011058	How To Maximize Your Profits Trading Options	https://www.udemy.com/how-to-maximize-your-pro...	True	200	1276	45	26	Intermediate Level	

In [4]:
```
data.shape
```

Out[4]:
```
(3678, 12)
```

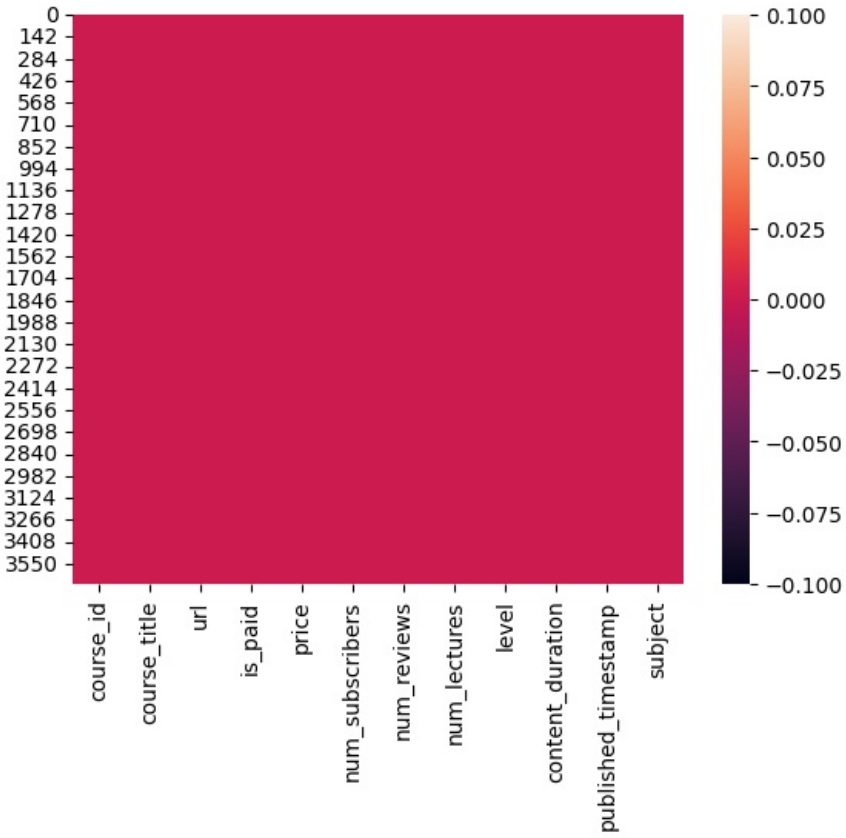
In [5]:
```
data.info()
```

```
<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
```

2. Show null data on Dataset

In [6]:
```
sns.heatmap(data.isnull())
```

Out[6]:
```
<Axes: >
```



3. Drop the duplicate data

In [7]:
```
dup=data.duplicated().any()
dup
```

Out[7]:
```
True
```

In [8]:
```
data=data.drop_duplicates()
dup=data.duplicated().any()
dup
```

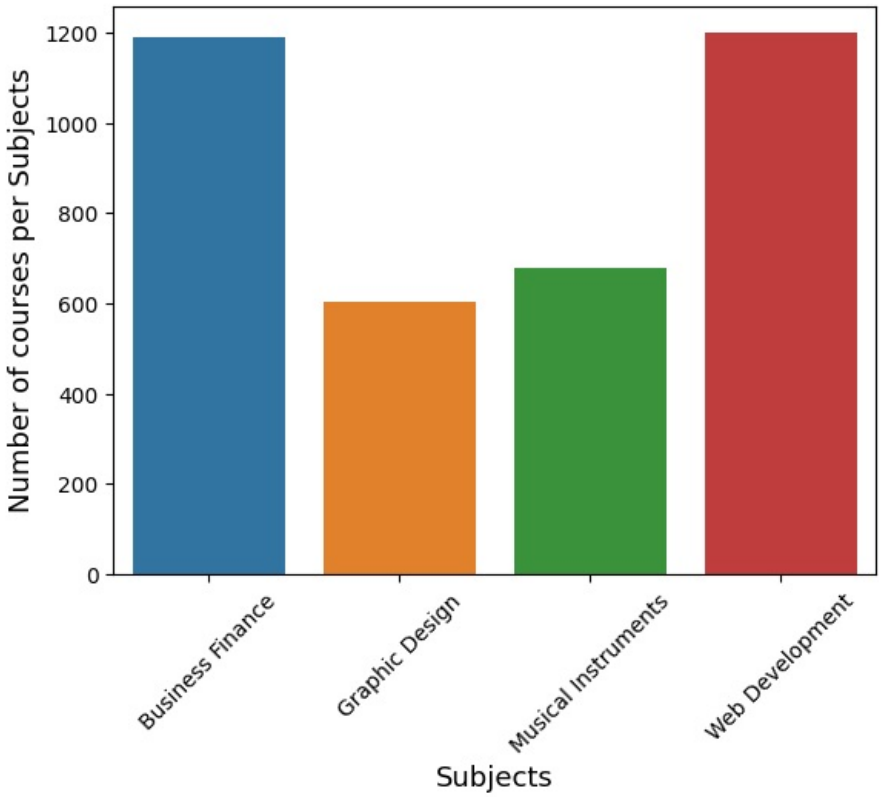
Out[8]:
```
False
```

In [9]:
```
data['subject'].value_counts()
```

Out[9]:
```
Web Development        1199
Business Finance       1191
Musical Instruments     680
Graphic Design          602
Name: subject, dtype: int64
```

4. Show number courses per subjects

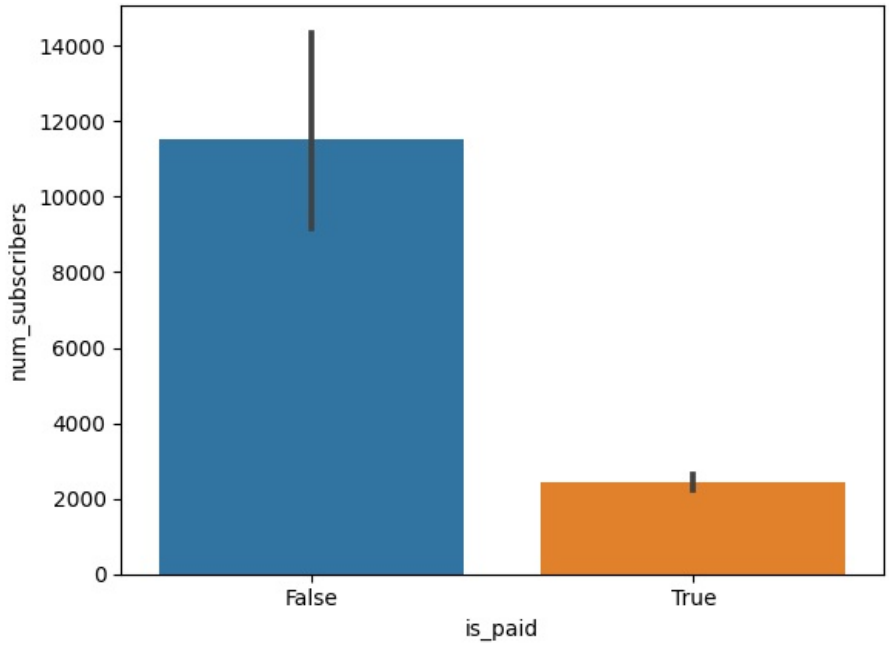
In [10]:
```
sns.countplot(x=data['subject'])
plt.xlabel('Subjects',fontsize=13)
plt.ylabel('Number of courses per Subjects',fontsize=13)
plt.xticks(rotation=45)
plt.show()
```



5. Show the number of subscribers of paid & free courses

In [12]:
```
sns.barplot(x='is_paid',y='num_subscribers',data=data)
```

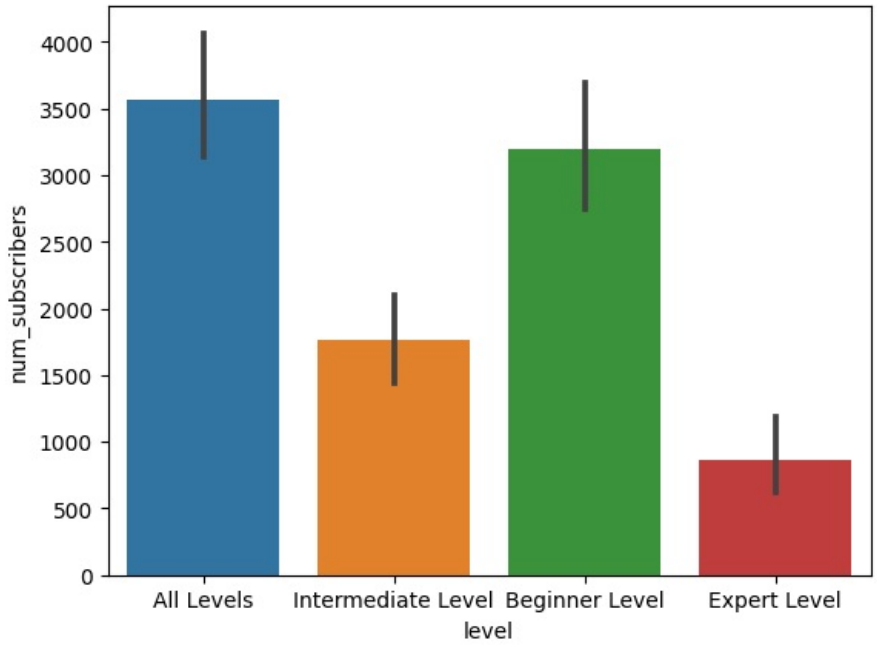
Out[12]:
```
<Axes: xlabel='is_paid', ylabel='num_subscribers'>
```



6. Show the number of subscribers of Levels

In [13]:
```
sns.barplot(x='level',y='num_subscribers',data=data)
```

Out[13]:
```
<Axes: xlabel='level', ylabel='num_subscribers'>
```



In [15]:
```
data.columns
```

Out[15]:
```
Index(['course_id', 'course_title', 'url', 'is_paid', 'price',
       'num_subscribers', 'num_reviews', 'num_lectures', 'level',
       'content_duration', 'published_timestamp', 'subject'],
      dtype='object')
```

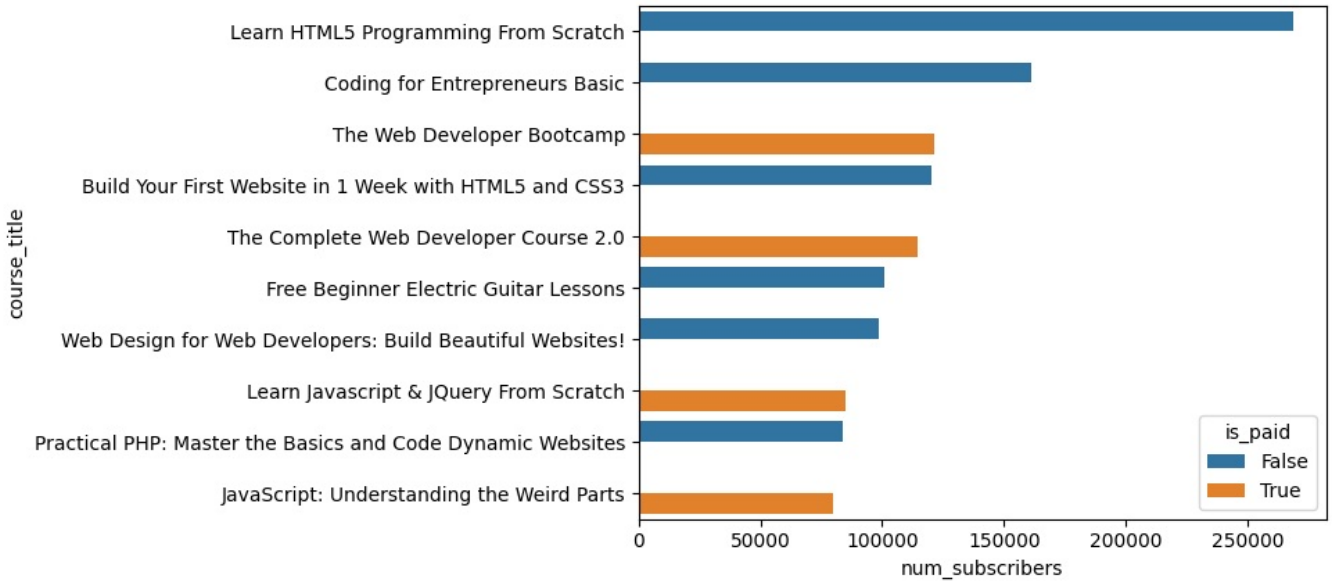
7. Show top 10 course titles based on number of subscribers

In [23]:
```
data[data['num_subscribers'].max()==data['num_subscribers']]
```

In [24]:
```
top_10=data.sort_values(by='num_subscribers',ascending=False).head(10)
```

In [26]:
```
sns.barplot(x='num_subscribers',y='course_title',data=top_10,hue='is_paid')
```

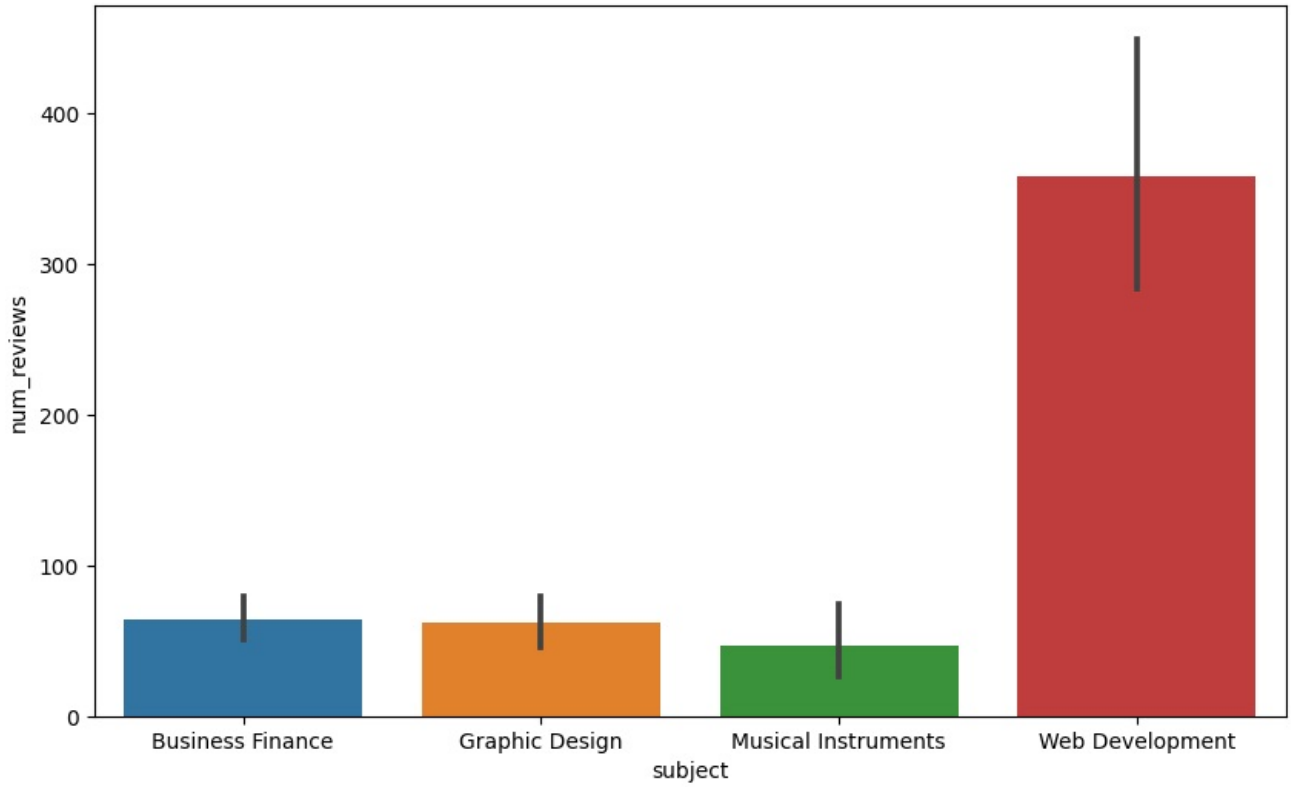
Out[26]:
```
<Axes: xlabel='num_subscribers', ylabel='course_title'>
```



8. Show num of reviews based on Subjects

In [31]:
```
plt.figure(figsize=(10,6))
sns.barplot(x='subject',y='num_reviews',data=data)
```

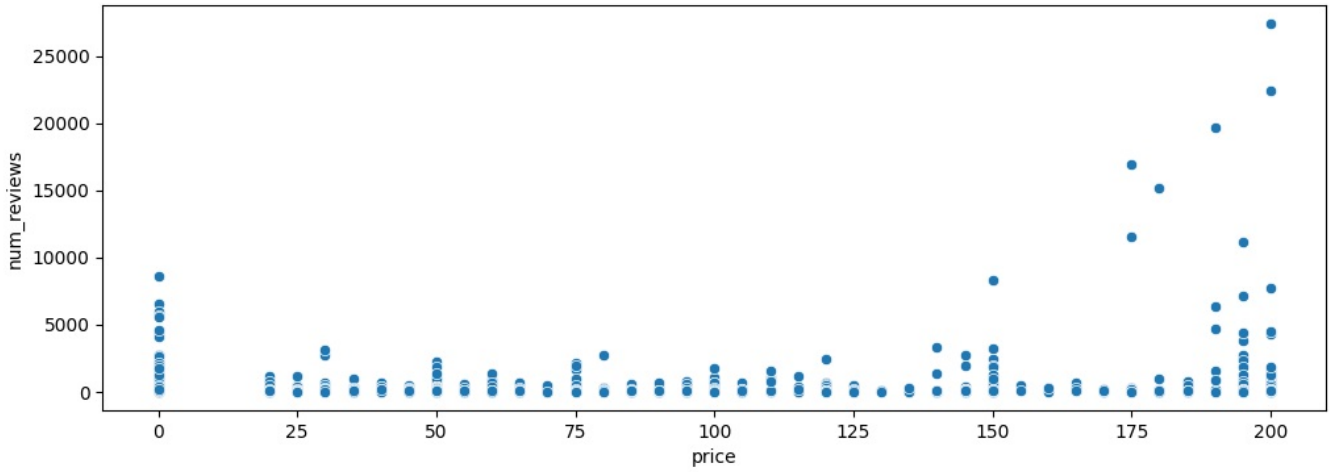
Out[31]:
```
<Axes: xlabel='subject', ylabel='num_reviews'>
```



9. Plotting price vs number of reviews (& number of subscribers)

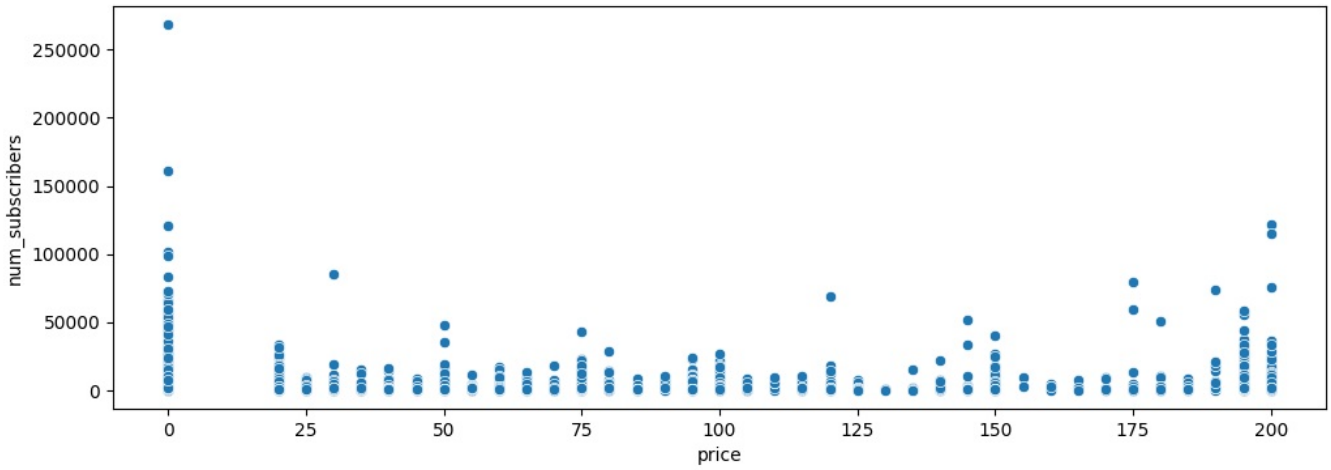
In [32]:
```
plt.figure(figsize=(12,4))
sns.scatterplot(x='price',y='num_reviews',data=data)
```

Out[32]:
```
<Axes: xlabel='price', ylabel='num_reviews'>
```



In [33]:
```
plt.figure(figsize=(12,4))
sns.scatterplot(x='price',y='num_subscribers',data=data)
```

Out[33]:
```
<Axes: xlabel='price', ylabel='num_subscribers'>
```



10. Top 5 python courses

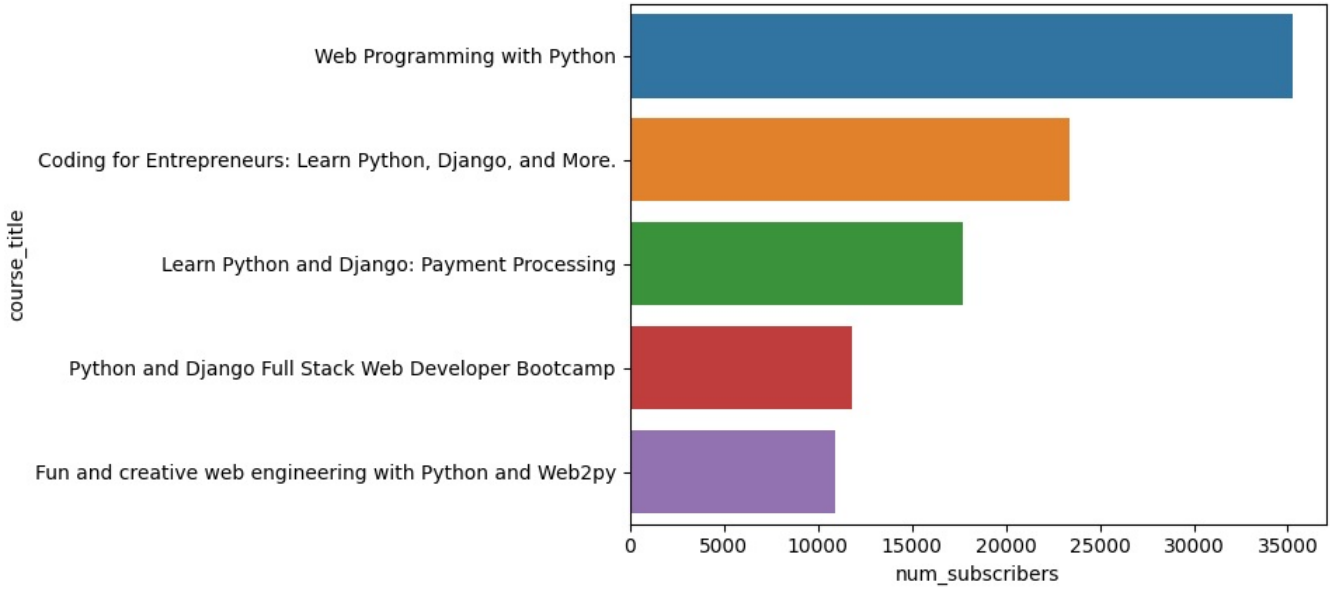
In [38]:
```
len(data[data['course_title'].str.contains('python',case=False)])
```

Out[38]: 29

In [47]:
```
python=data[data['course_title'].str.contains('python',case=False)]. \
sort_values('num_subscribers',ascending=False).head(5)
```

In [48]:
```
sns.barplot(x='num_subscribers',y='course_title',data=python)
```

Out[48]: <Axes: xlabel='num_subscribers', ylabel='course_title'>



In [50]:
```
data['Year']=data['published_timestamp'].dt.year
```

In [51]:
```
data.head(2)
```

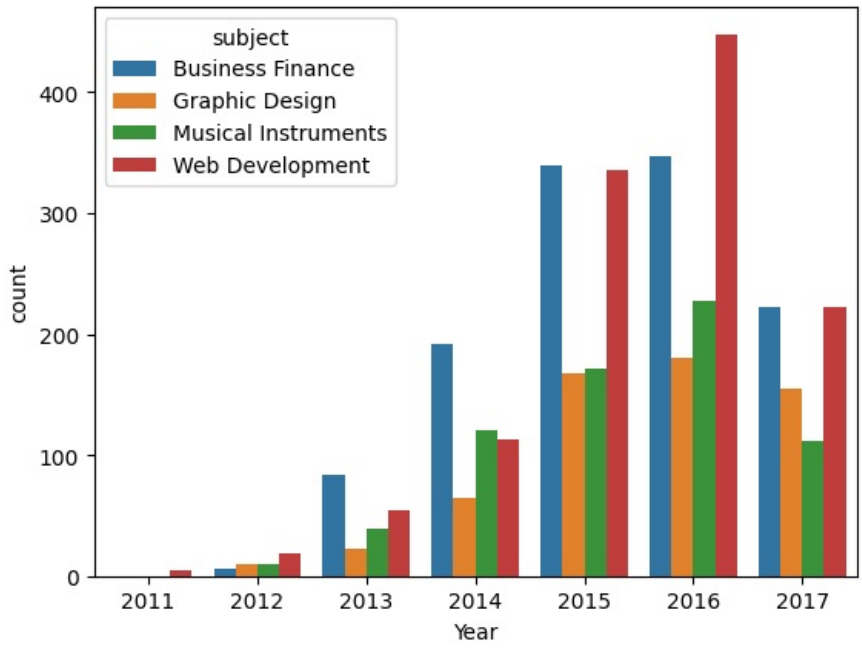
Out[51]:

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_durati
0	1070968	Ultimate Investment Banking Course	https://www.udemy.com/ultimate-investment-bank...	True	200	2147	23	51	All Levels	
1	1113822	Complete GST Course & Certification - Grow You...	https://www.udemy.com/goods-and-services-tax/	True	75	2792	923	274	All Levels	3

11. Subjects uploaded yearly

In [54]:
```
sns.countplot(x='Year',data=data,hue='subject')
```

Out[54]:
```
<Axes: xlabel='Year', ylabel='count'>
```



In [55]:
```
data.groupby('Year')['subject'].value_counts()
```

Out[55]:
```
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
```

In [ ]: