```
pip install opendatasets
```

```
Collecting opendatasets
  Downloading opendatasets-0.1.22-py3-none-any.whl (15 kB)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-
packages (from opendatasets) (4.66.4)
Requirement already satisfied: kaggle in
/usr/local/lib/python3.10/dist-packages (from opendatasets) (1.6.14)
Requirement already satisfied: click in
/usr/local/lib/python3.10/dist-packages (from opendatasets) (8.1.7)
Requirement already satisfied: six>=1.10 in
/usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets)
(1.16.0)
Requirement already satisfied: certifi>=2023.7.22 in
/usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets)
(2024.2.2)
Requirement already satisfied: python-dateutil in
/usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets)
(2.8.2)
Requirement already satisfied: requests in
/usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets)
(2.31.0)
Requirement already satisfied: python-slugify in
/usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets)
(8.0.4)
Requirement already satisfied: urllib3 in
/usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets)
(2.0.7)
Requirement already satisfied: bleach in
/usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets)
(6.1.0)
Requirement already satisfied: webencodings in
/usr/local/lib/python3.10/dist-packages (from bleach->kaggle-
>opendatasets) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in
/usr/local/lib/python3.10/dist-packages (from python-slugify->kaggle-
>opendatasets) (1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->kaggle-
>opendatasets) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests->kaggle-
>opendatasets) (3.7)
Installing collected packages: opendatasets
Successfully installed opendatasets-0.1.22
```

```python
import opendatasets as od
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings                                    #to filter and ignore warning
messages
warnings.filterwarnings('ignore')

od.download("https://www.kaggle.com/datasets/andrewmvd/udemy-courses/
code")
```

```
Please provide your Kaggle credentials to download this dataset. Learn
more: http://bit.ly/kaggle-creds
Your Kaggle username: ":"muhammadabdulumair
Your Kaggle Key: ··········
Dataset URL: https://www.kaggle.com/datasets/andrewmvd/udemy-courses
Downloading udemy-courses.zip to ./udemy-courses

100%|██████████| 200k/200k [00:00<00:00, 45.4MB/s]
```

```
import pandas as pd

df = pd.read_csv("/content/udemy-courses/udemy_courses.csv")

df
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 3678,\n  \"fields\":
[\n    {\n      \"column\": \"course_id\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 343273,\n        \"min\":
8324,\n        \"max\": 1282064,\n        \"num_unique_values\":
3672,\n        \"samples\": [\n          26648,\n          1121580,\n
1076222\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"course_title\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 3663,\n
\"samples\": [\n          \"Photoshop - Automatiza\\u00e7\\u00e3o com
Adobe Script\",\n          \"Forex MetaTrader 4: Master MT4 Like A Pro
Forex Trader\",\n          \"* An Integrated Approach to the
Fundamentals of Accounting\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"url\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 3672,\n
\"samples\": [\n          \"https://www.udemy.com/how-to-play-guitar-
really-understand-music/\",\n
\"https://www.udemy.com/wordpress-website-for-beginners/\",\n
\"https://www.udemy.com/the-most-popular-techniques-in-photoshop/\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"is_paid\",\n
\"properties\": {\n        \"dtype\": \"boolean\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n          false,\n
true\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"price\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 61,\n        \"min\": 0,\n        \"max\": 200,\n
\"num_unique_values\": 38,\n        \"samples\": [\n          130,\n
110\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"num_subscribers\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 9504,\n        \"min\": 0,\n
\"max\": 268923,\n        \"num_unique_values\": 2197,\n
\"samples\": [\n          136,\n          251\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"num_reviews\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
935,\n        \"min\": 0,\n        \"max\": 27445,\n
\"num_unique_values\": 511,\n        \"samples\": [\n          265,\n
66\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"num_lectures\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 50,\n        \"min\": 0,\n
\"max\": 779,\n        \"num_unique_values\": 229,\n
\"samples\": [\n          342,\n          34\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"level\",\n      \"properties\": {\
n        \"dtype\": \"category\",\n        \"num_unique_values\": 4,\n
\"samples\": [\n          \"Intermediate Level\",\n          \"Expert
Level\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"content_duration\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 6.053840414790038,\n        \"min\":
0.0,\n        \"max\": 78.5,\n        \"num_unique_values\": 105,\n
\"samples\": [\n          46.5,\n          70.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"published_timestamp\",\n
\"properties\": {\n        \"dtype\": \"object\",\n
\"num_unique_values\": 3672,\n        \"samples\": [\n
\"2012-10-13T23:40:19Z\",\n          \"2017-02-26T18:29:53Z\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"subject\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 4,\n        \"samples\": [\n          \"Graphic
Design\",\n          \"Web Development\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

```
df.dtypes
```

```
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
```

```
df.head(10)
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 3678,\n  \"fields\":
[\n    {\n      \"column\": \"course_id\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 343273,\n        \"min\":
8324,\n        \"max\": 1282064,\n        \"num_unique_values\":
3672,\n        \"samples\": [\n          26648,\n          1121580,\n
1076222\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"course_title\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 3663,\n
\"samples\": [\n          \"Photoshop - Automatiza\\u00e7\\u00e3o com
Adobe Script\",\n          \"Forex MetaTrader 4: Master MT4 Like A Pro
Forex Trader\",\n          \"* An Integrated Approach to the
Fundamentals of Accounting\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"url\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 3672,\n
\"samples\": [\n          \"https://www.udemy.com/how-to-play-guitar-
really-understand-music/\",\n
\"https://www.udemy.com/wordpress-website-for-beginners/\",\n
\"https://www.udemy.com/the-most-popular-techniques-in-photoshop/\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"is_paid\",\n
\"properties\": {\n        \"dtype\": \"boolean\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n          false,\n
true\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"price\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 61,\n        \"min\": 0,\n        \"max\": 200,\n
\"num_unique_values\": 38,\n        \"samples\": [\n          130,\n
110\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"num_subscribers\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 9504,\n        \"min\": 0,\n
\"max\": 268923,\n        \"num_unique_values\": 2197,\n
\"samples\": [\n          136,\n          251\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"num_reviews\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
```

```
935,\n        \"min\": 0,\n        \"max\": 27445,\n
\"num_unique_values\": 511,\n        \"samples\": [\n          265,\n
66\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"num_lectures\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 50,\n        \"min\": 0,\n
\"max\": 779,\n        \"num_unique_values\": 229,\n
\"samples\": [\n          342,\n          34\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"level\",\n      \"properties\": {\
n        \"dtype\": \"category\",\n        \"num_unique_values\": 4,\n
\"samples\": [\n          \"Intermediate Level\",\n          \"Expert
Level\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"content_duration\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 6.053840414790038,\n        \"min\":
0.0,\n        \"max\": 78.5,\n        \"num_unique_values\": 105,\n
\"samples\": [\n          46.5,\n          70.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"published_timestamp\",\n
\"properties\": {\n        \"dtype\": \"object\",\n
\"num_unique_values\": 3672,\n        \"samples\": [\n
\"2012-10-13T23:40:19Z\",\n          \"2017-02-26T18:29:53Z\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"subject\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 4,\n        \"samples\": [\n          \"Graphic
Design\",\n          \"Web Development\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

```
df.shape
```

```
(3678, 12)
```

```
print("The number of rows :",df.shape[0])
print("The number of columns:",df.shape[1])
```

```
The number of rows : 3678
The number of columns: 12
```

```
df.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   object
 11  subject              3678 non-null   object
dtypes: bool(1), float64(1), int64(5), object(5)
memory usage: 319.8+ KB
```

```
print("Is there any null value in the
dataset?",df.isnull().sum().any())
```

```
Is there any null value in the dataset? False
```

```
df.isnull().sum()
```

```
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
```

```
print("Is there any duplicates value ?",df.duplicated().any())
```

```
Is there any duplicates value ? True
```

```
df.drop_duplicates(inplace =True)
print("Is there any duplicates value ?",df.duplicated().any())
```

```
Is there any duplicates value ? False
```

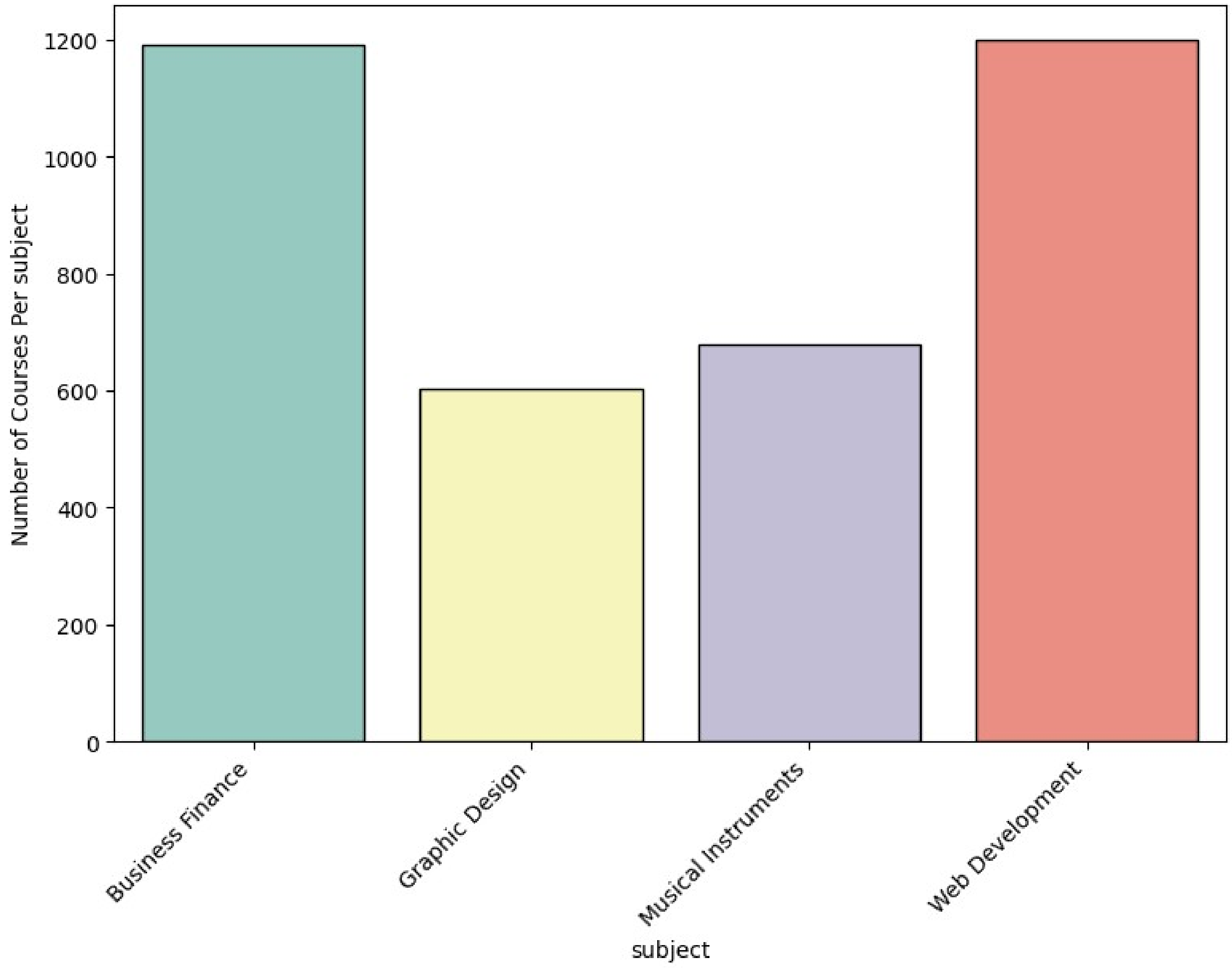
```
df.columns
```

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

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

```
subject
Web Development        1199
Business Finance       1191
Musical Instruments     680
Graphic Design          602
Name: count, dtype: int64
```

```python
plt.figure(figsize = (9,6))

#Create the countplot
sns.countplot(x = 'subject',data =df,palette ='Set3',edgecolor ='black')

#add labels
plt.ylabel("Number of Courses Per subject")
# Customize x-axis ticks
plt.xticks(rotation =45 ,ha ='right')  # Rotate x-axis labels for better readability

#show the plot
plt.show()
```



```
df.columns
```

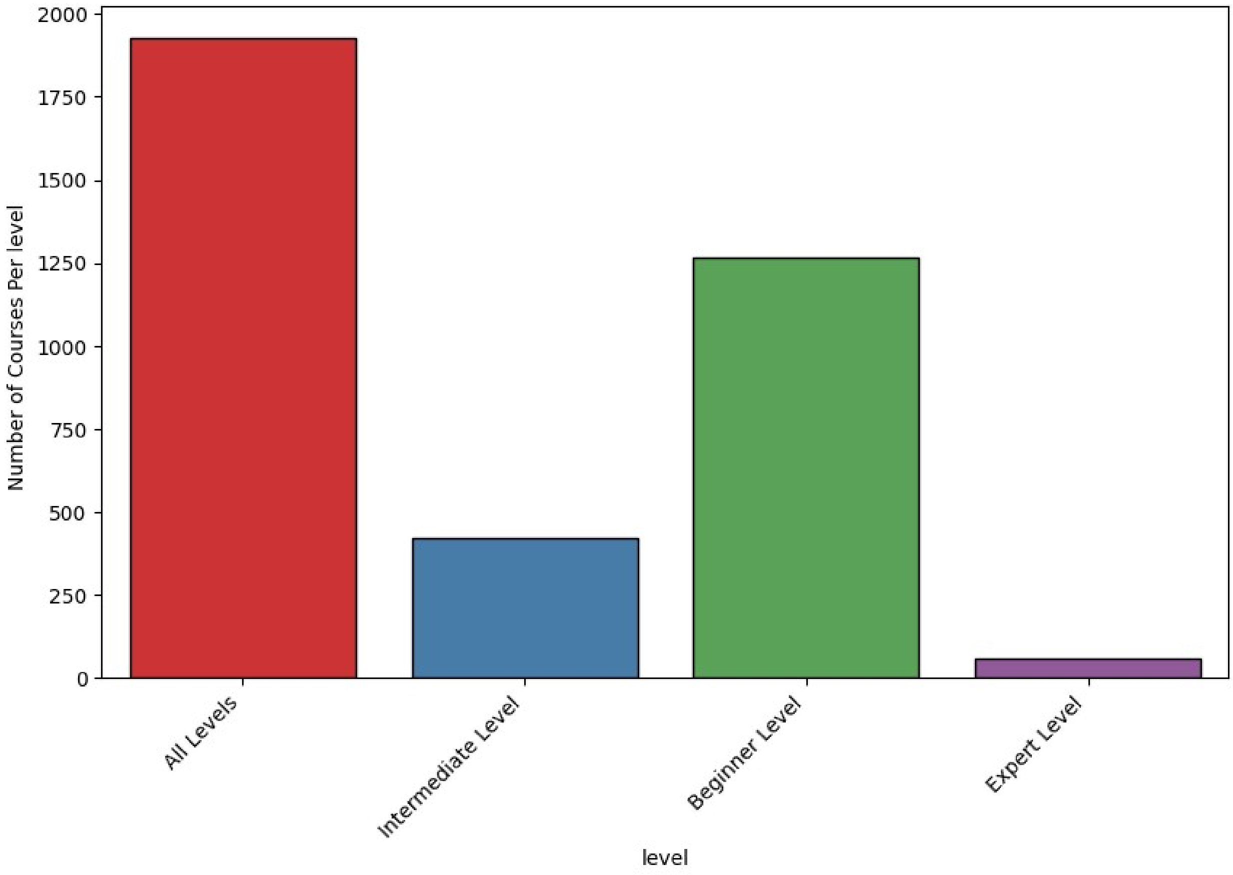
```
Index(['course_id', 'course_title', 'url', 'is_paid', 'price',
       'num_subscribers', 'num_reviews', 'num_lectures', 'level',
       'content_duration', 'published_timestamp', 'subject'],
      dtype='object')
```

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

```
level
All Levels            1925
Beginner Level        1268
Intermediate Level     421
Expert Level            58
Name: count, dtype: int64
```

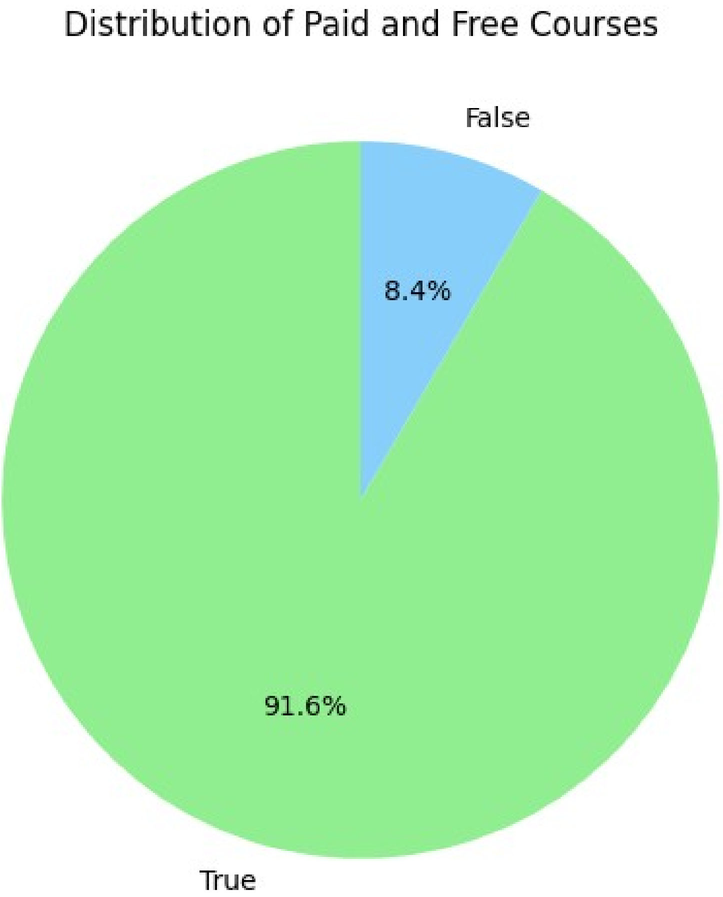
```
plt.figure(figsize = (10,6))

#Create the countplot
sns.countplot(x = 'level',data =df,palette ='Set1',edgecolor ='black')
```

```
#add labels
plt.ylabel("Number of Courses Per level ")

# Customize x-axis ticks
plt.xticks(rotation =45 ,ha ='right')  # Rotate x-axis labels for
better readability

#show the plot
plt.show()
```



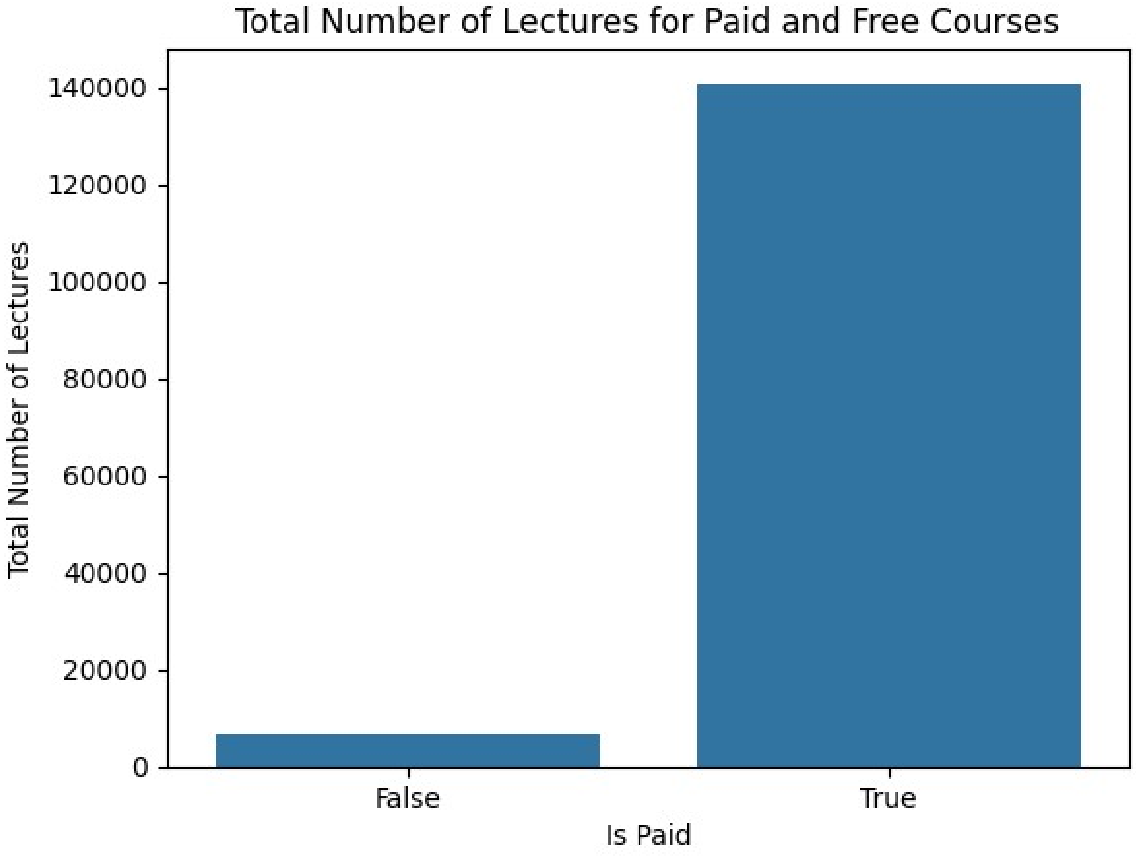
```
df.columns

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

df['is_paid'].value_counts()

is_paid
True     3362
False     310
Name: count, dtype: int64
```

```
value_counts = df['is_paid'].value_counts()

# Plotting pie chart
plt.figure(figsize=(6, 6))
plt.pie(value_counts, labels=value_counts.index, autopct='%1.1f%%',
startangle=90, colors=['lightgreen', 'lightskyblue'])
plt.title('Distribution of Paid and Free Courses')
plt.show()
```



```
df.groupby(['is_paid'])['num_lectures'].sum()

is_paid
False      6639
True     140756
Name: num_lectures, dtype: int64

grouped_data = df.groupby(['is_paid'])
['num_lectures'].sum().reset_index()

# Plotting bar plot
```

```
sns.barplot(x='is_paid', y='num_lectures', data=grouped_data)
plt.xlabel('Is Paid')
plt.ylabel('Total Number of Lectures')
plt.title('Total Number of Lectures for Paid and Free Courses')
plt.show()
```



```
df.columns
```

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

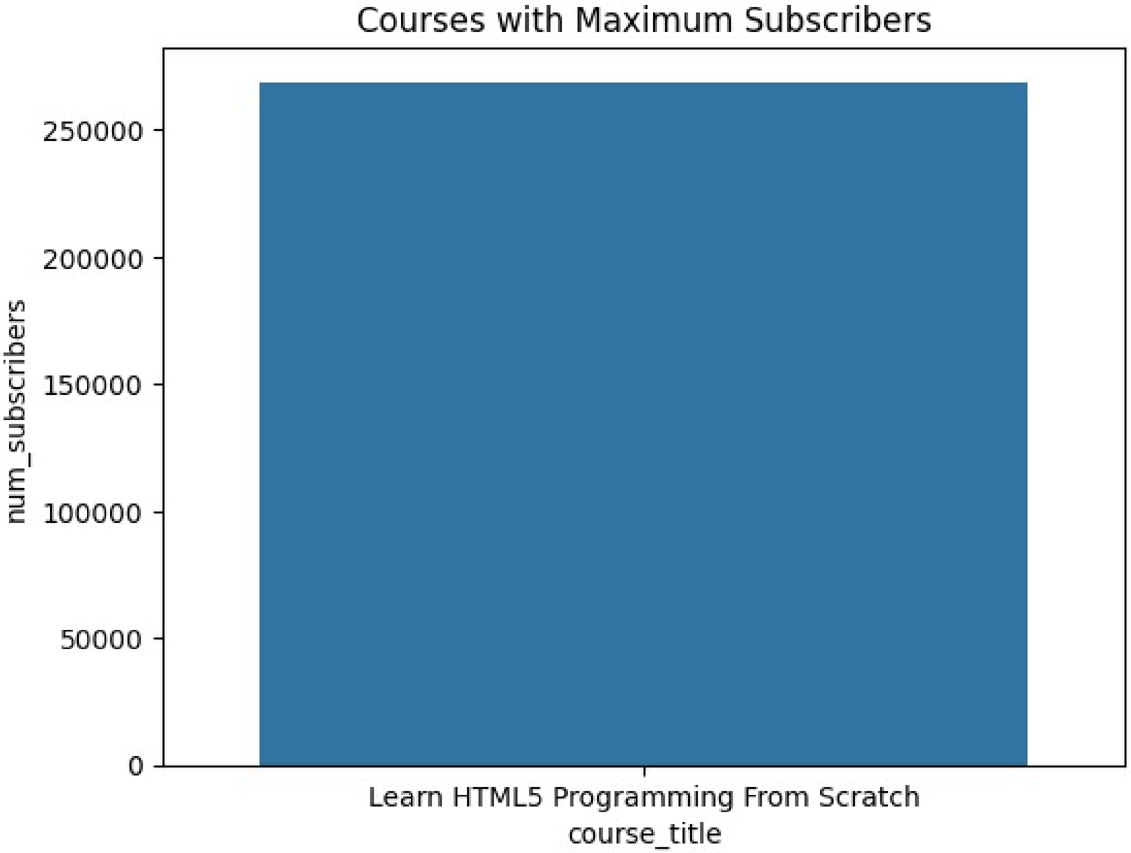
```
Max_data = df['num_subscribers'].max() == df['num_subscribers']
```

```
df[Max_data]['course_title']
```

```
2827    Learn HTML5 Programming From Scratch
Name: course_title, dtype: object
```

```
sns.barplot(x ='course_title',y ='num_subscribers',data= df[Max_data])
```

```
plt.title('Courses with Maximum Subscribers')
plt.show()
```


Courses with Maximum Subscribers
num_subscribers
Learn HTML5 Programming From Scratch
course_title