Udemy Courses data Analysis

image.png

This project is to perform Exploratory Data Analysis on the dataset contains detailed information on all available Udemy courses on Oct 10,2022.The information of over 209k courses was collected by web scraping the Udemy website. Udemy holds 209,734 courses and 73,514 instructors teaching courses in 79 languages in 13 different categories.

- **Link to the Dataset used- Source**
  The dataset contains information of all the Udemy courses,their prices,rates,rating,subscribers,average rating,number of reviews,number of comments,number of lectures,etc.

The Libraries used in the project are:

- Matplotlib(for visualization of data) Explore
- Seaborn (used alongside Matplotlib for visualization) Explore
- Numpy (used for operations on numeric data) Explore
- Pandas (used for utilising DataFrames and organising the data)Explore
- Jovian (used for downloading dataset and to run, save and upload the Notebook)Explore

To install all required libraries,run the following Command:
pip install matplotlib seaborn numpy pandas jovian --upgrade

The following Tasks are implemented in the Project:

- Data Preparation and Cleaning
- Exploratory Analysis and Visualization
- Asking and Answering Questions
- Inferences and Conclusion
- References and Future Work

LET'S DIVE INTO THE PROJECT !!!

Downloading Dataset

There are several options for getting the dataset into Jupyter:

- Download the CSV manually and upload it via Jupyter's GUI
- Use the urlretrieve function from the urllib.request to download CSV files from a raw URL
- Use a helper library, e.g., opendatasets, which contains a collection of curated datasets and provides a helper function for direct download.

We'll use the opendatasets helper library to download the files.

```
import opendatasets as od
```

```
---------------------------------------------------------------------------
```

```
ModuleNotFoundError                       Traceback (most recent call last)
/tmp/ipykernel_45/4147301787.py in <module>
----> 1 import opendatasets as od

ModuleNotFoundError: No module named 'opendatasets'
```

You may get an error like
ModuleNotFoundError: No module named 'opendatasets'
So you have to first install the opendatasets in your Current
working Notebook

```
#Here we are installing the opendatasets in our current notebook
!pip install opendatasets --upgrade --quiet
#we use quiet so that output of installed functions will not visible.
```

```
dataset_url='https://www.kaggle.com/datasets/hossaingh/udemy-courses?select=Course_info
```

Let's begin by downloading the dataset

```
import opendatasets as od
od.download(dataset_url)
```

```
Please provide your Kaggle credentials to download this dataset. Learn more:
http://bit.ly/kaggle-creds
Your Kaggle username:
```

```
import os
```

```
data_dir='./udemy-courses'
```

Let's look through our files in directory.

```
os.listdir(data_dir)
```

```
['Course_info.csv', 'Comments.csv']
```

Data Preparation and Cleaning

The raw data is now obtained.First we need to clean and simplify the data in order to prepare it for Analysis.

The .csv file which we downloaded from Kaggle will convert to a Pandas DataFrame and clean to extract only the columns which will be needed for analysis.

Let us import the pandas to read csv file and for plotting import matplotlib seaborn and their modules.

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

sns.set_style('darkgrid')
```

Let's read our file using pandas.

```
udemy_courses_data=pd.read_csv('./udemy-courses/Course_info.csv')
```

Taking sample of our data:

```
#I always like to get rough knowlwdge about our data by seeing some random rows.
udemy_courses_data.sample(10)
```

	id	title	is_paid	price	headline	num_subscribers	avg_rating	num_reviews	num_com
59740	2179498.0	The eCom Brand Accelerator	True	19.99	A Proven eCom Expert Shares How To Build A Bra...	19.0	4.250000	2.0	
15904	778462.0	Cerveja Artesanal	True	179.90	Aprenda a fazer cerveja em casa	984.0	3.900000	342.0	
51744	1931324.0	Mechanical Engineering and Electrical Engineer...	True	29.99	Learn how things work! Boilers, engines, valve...	14001.0	4.716667	1054.0	
30074	1312016.0	Animação do Hand Lettering	True	34.99	aprenda a animar a sua arte no Photoshop	581.0	4.950000	142.0	
38403	1546264.0	How to Make a Feature Film with No Money and N...	True	39.99	An uplifting boost for your filmmaking self-co...	111.0	4.000000	9.0	
68022	2397186.0	Mastering Palo Alto Networks	True	124.99	Dominate and take control of all the features ...	8191.0	4.272728	1693.0	
103667	3264692.0	The Packed Calendar Photography Business Bluep...	True	19.99	Consistently Book Photo Shoots Every Week Usin...	2.0	0.000000	0.0	
102452	3236791.0	Tips Menerapkan Performance Appraisal yang Sesuai	True	349.00	Tips Menerapkan Performance Appraisal yang Sesuai	20.0	4.857143	7.0	
7954	440382.0	Sıfırdan Linux Ağ ve Sistem Yöneticiliği Eğitimi	True	249.99	Zengin içerikli ve uygulamalı eğitimle sizde ...	1305.0	3.600000	262.0	

	id	title	is_paid	price	headline	num_subscribers	avg_rating	num_reviews	num_com
192988	4662378.0	Aprenda SQL Básico	False	0.00	SQL Básico	439.0	3.833333	34.0	

10 rows × 27 columns

A first look at the data

Let's do some experiment with our data ,try to clean some unwanted data.

```
print(f'The df has {udemy_courses_data.shape[0]} rows and {udemy_courses_data.shape[1]}

print(f"The df contains data on {udemy_courses_data['id'].nunique()} unique courses")

if udemy_courses_data['id'].nunique() == udemy_courses_data.shape[0]:
    print("There's exactly one row per course.")
else:
    print("There's not exactly one row per course.")

# So, on 10 Oct 2022, there was more than 2 lakh courses on Udemy!
```

```
The df has 209734 rows and 20 columns.
The df contains data on 209734 unique courses
There's exactly one row per course.
```

Let's take some more information about our data

```
udemy_courses_data.info()

# we can clearly see that our data is in float,bool,object.
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209734 entries, 0 to 209733
Data columns (total 20 columns):
 #   Column              Non-Null Count   Dtype  
---  ------              --------------   -----  
 0   id                  209734 non-null  float64
 1   title               209734 non-null  object 
 2   is_paid             209734 non-null  bool   
 3   price               209734 non-null  float64
 4   headline            209707 non-null  object 
 5   num_subscribers     209734 non-null  float64
 6   avg_rating          209734 non-null  float64
 7   num_reviews         209734 non-null  float64
 8   num_comments        209734 non-null  float64
 9   num_lectures        209734 non-null  float64
 10  content_length_min  209734 non-null  float64
 11  published_time      209734 non-null  object 
 12  last_update_date    209597 non-null  object
 13  category            209734 non-null  object
 14  subcategory         209734 non-null  object
 15  topic               208776 non-null  object
 16  language            209734 non-null  object
 17  course_url          209734 non-null  object
 18  instructor_name     209729 non-null  object
 19  instructor_url      209307 non-null  object
dtypes: bool(1), float64(8), object(11)
memory usage: 30.6+ MB
```

In our data the columns "published_time" and "last_update_date" data type should be in datetime format but here it is in object format so we need to change it to datetime format.

```
udemy_courses_data['published_time']=pd.to_datetime(udemy_courses_data['published_time'
udemy_courses_data['last_update_date']=pd.to_datetime(udemy_courses_data['last_update_d

#Now when we check the our data again
udemy_courses_data.info()
#Those columns are in datetime format
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209734 entries, 0 to 209733
Data columns (total 20 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   id                  209734 non-null  float64
 1   title               209734 non-null  object
 2   is_paid             209734 non-null  bool
 3   price               209734 non-null  float64
 4   headline            209707 non-null  object
 5   num_subscribers     209734 non-null  float64
 6   avg_rating          209734 non-null  float64
 7   num_reviews         209734 non-null  float64
 8   num_comments        209734 non-null  float64
 9   num_lectures        209734 non-null  float64
 10  content_length_min  209734 non-null  float64
 11  published_time      209734 non-null  datetime64[ns, UTC]
 12  last_update_date    209597 non-null  datetime64[ns]
 13  category            209734 non-null  object
 14  subcategory         209734 non-null  object
 15  topic               208776 non-null  object
 16  language            209734 non-null  object
 17  course_url          209734 non-null  object
 18  instructor_name     209729 non-null  object
 19  instructor_url      209307 non-null  object
dtypes: bool(1), datetime64[ns, UTC](1), datetime64[ns](1), float64(8), object(9)
memory usage: 30.6+ MB
```

To get the information about our numerical columns we can also use describe function

```
udemy_courses_data.describe()
```

	id	price	num_subscribers	avg_rating	num_reviews	num_comments	num_lec
count	2.097340e+05	209734.000000	2.097340e+05	209734.000000	209734.000000	209734.000000	209734.00
mean	3.015403e+06	81.665529	3.096992e+03	3.747179	244.358812	44.874589	36.54
std	1.342558e+06	117.317846	1.558132e+04	1.533711	2458.098276	355.773107	51.87
min	1.769000e+03	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.00
25%	1.950734e+06	19.990000	2.600000e+01	3.800000	3.000000	1.000000	11.00
50%	3.292863e+06	34.990000	2.060000e+02	4.333334	17.000000	5.000000	22.00
75%	4.189458e+06	99.990000	1.435000e+03	4.625000	74.000000	18.000000	42.00
max	4.914146e+06	999.990000	1.752364e+06	5.000000	436457.000000	39040.000000	1095.00

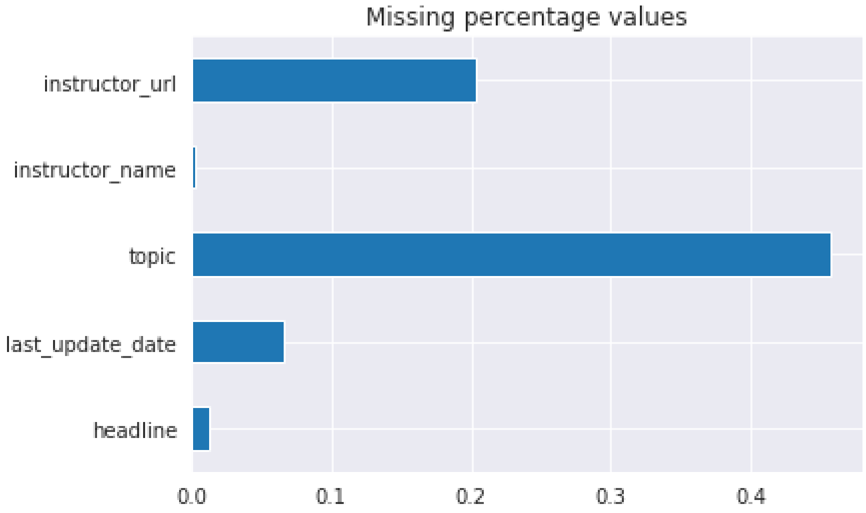
Before analysing the data let's check percentage of missing NA values in each column in our data.

```
missing_percentage=udemy_courses_data.isna().sum()/len(udemy_courses_data)*100
missing_percentage
```

```
id                   0.000000
title                0.000000
is_paid              0.000000
price                0.000000
headline             0.012873
num_subscribers      0.000000
avg_rating           0.000000
num_reviews          0.000000
num_comments         0.000000
num_lectures         0.000000
content_length_min   0.000000
published_time       0.000000
last_update_date     0.065321
category             0.000000
subcategory          0.000000
topic                0.456769
language             0.000000
course_url           0.000000
instructor_name      0.002384
instructor_url       0.203591
dtype: float64
```

```
#We can also calculate total missing data
print(f"Out of 210k courses, only {round(missing_percentage.sum(),2)}% of data is missi
print('-----------------------------------------------------------')
#We can check only non zero values data and also draw the graph of that
missing_percentage[missing_percentage!=0].plot(kind='barh')
plt.title('Missing percentage values');

#Most missing values in "topic" column.
```

```
Out of 210k courses, only 0.74% of data is missing.
-----------------------------------------------------------
```



From th above plot we can see that in topic column there is percentage of data missing around 0.45% which is greater than half of the total missing percentage.And then from column of instructor_url.

To get a sense of the timeframe of our data, let's see what's the "oldest" data point that we have.

Let's also see the publication date of the "youngest" course, and the most recent last_update_date that we have.

```
udemy_courses_data.published_time.describe()
```

```
/tmp/ipykernel_39/1368805534.py:1: FutureWarning: Treating datetime data as categorical rather than numeric in `.describe` is deprecated and will be removed in a future version of pandas. Specify `datetime_is_numeric=True` to silence this warning and adopt the future behavior now.
  udemy_courses_data.published_time.describe()
```

```
count                        209734
unique                       209562
top       2016-08-06 07:33:16+00:00
freq                              4
first     2010-01-12 18:09:46+00:00
last      2022-10-05 17:04:08+00:00
Name: published_time, dtype: object
```

In above cell we get an warning because our data type is in datetime and we know according to the definition of describe() (The describe() method returns description of the data in theDataFrame. If the DataFrame contains numerical data, the description contains these information for each column: count) format to avoid we can convert this data time into numeric(For this only) to avoid warning.

```
udemy_courses_data.published_time.describe(datetime_is_numeric=True)

#So, the oldest course is published on 2010-01-12(12 January 2010).
#It's a bit strange because the Udemy was officcially founded on May 11,2010(source: Wi
#let's check
```

```
count                                  209734
mean     2020-02-06 16:13:05.572520704+00:00
min                2010-01-12 18:09:46+00:00
25%      2018-12-08 20:58:21.750000128+00:00
50%                2020-08-21 02:41:58+00:00
75%         2021-08-30 21:35:41.500000+00:00
max                2022-10-05 17:04:08+00:00
Name: published_time, dtype: object
```

Let's quickly zoom in on courses published before May 11, 2010

```
print(f"There are only {len(udemy_courses_data[udemy_courses_data.published_time<'2010-
```

```
There are only 2 such courses
```

```
#let's see those 2 courses
udemy_courses_data[udemy_courses_data.published_time<'2010-05-11']
```

	id	title	is_paid	price	headline	num_subscribers	avg_rating	num_reviews	num_comments	nur
1	1769.0	The Lean Startup Talk at Stanford E-Corner	False	0.00	Debunking Myths of Entrepreneurship A startup ...	26474.0	4.50	709.0	112.0	
6	2762.0	Simple Strategy for Swing Trading the Stock Ma...	True	39.99	Use my favorite Technical Indicator and the Tr...	3309.0	3.85	958.0	241.0	

Mystery solved(source:Wikipedia)

The site launched by Bali,Octay Caglar and Gagan Biyani in early May 2010.

let's also see the last updates of the courses

```
udemy_courses_data.last_update_date.describe(datetime_is_numeric=True)

#The most recent "last_update_date" is Oct 10, 2022.
#We can also see that 75% of the courses updates on and before March 17, 2022.
```

```
count                         209597
mean     2020-10-26 22:44:29.706150912
```

```
min                  2012-10-31 00:00:00
25%                  2020-02-19 00:00:00
50%                  2021-04-28 00:00:00
75%                  2022-03-17 00:00:00
max                  2022-10-10 00:00:00
Name: last_update_date, dtype: object
```

Exploratory Analysis and Visualization

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

Let's see which unique categories/subcategories we have.Let's assign a number to each category, and let's print each category and all of its subcategories...... in aplhabetical order Let's apply some basic formatting to make everything look nice :)

```
counter=1
for category in sorted(list(udemy_courses_data['category'].unique())):

    print(f"CATEGORY {counter}: {category}")
    print("SUBCATEGORIES:")

    ## Print each ot its subcategories
    ## I needed to use list() twice here - I'd get an array when I'd use it just once.
    for subcategory in sorted(list(udemy_courses_data
          [udemy_courses_data['category'] == category]['subcategory'].unique())):

        print(f"- {subcategory}")

    print('--------------------')

    counter += 1
```

```
CATEGORY 1: Business
SUBCATEGORIES:
- Business Analytics & Intelligence
- Business Law
- Business Strategy
- Communication
- E-Commerce
- Entrepreneurship
- Human Resources
- Industry
- Management
- Media
- Operations
- Other Business
- Project Management
```

- Real Estate
- Sales

CATEGORY 2: Design

SUBCATEGORIES:

- 3D & Animation
- Architectural Design
- Design Tools
- Fashion Design
- Game Design
- Graphic Design & Illustration
- Interior Design
- Other Design
- User Experience Design
- Web Design

CATEGORY 3: Development

SUBCATEGORIES:

- Data Science
- Database Design & Development
- Game Development
- Mobile Development
- No-Code Development
- Programming Languages
- Software Development Tools
- Software Engineering
- Software Testing
- Web Development

CATEGORY 4: Finance & Accounting

SUBCATEGORIES:

- Accounting & Bookkeeping
- Compliance
- Cryptocurrency & Blockchain
- Economics
- Finance
- Finance Cert & Exam Prep
- Financial Modeling & Analysis
- Investing & Trading
- Money Management Tools
- Other Finance & Accounting
- Taxes

CATEGORY 5: Health & Fitness

SUBCATEGORIES:

- Dance
- Fitness
- General Health
- Martial Arts & Self Defense
- Meditation
- Mental Health
- Nutrition & Diet
- Other Health & Fitness
- Safety & First Aid
- Sports
- Yoga

CATEGORY 6: IT & Software

SUBCATEGORIES:

- Hardware
- IT Certifications
- Network & Security
- Operating Systems & Servers
- Other IT & Software

CATEGORY 7: Lifestyle

SUBCATEGORIES:

- Arts & Crafts
- Beauty & Makeup
- Esoteric Practices
- Food & Beverage
- Gaming
- Home Improvement & Gardening
- Other Lifestyle
- Pet Care & Training
- Travel

CATEGORY 8: Marketing

SUBCATEGORIES:

- Affiliate Marketing
- Branding
- Content Marketing
- Digital Marketing
- Growth Hacking
- Marketing Analytics & Automation
- Marketing Fundamentals

- Other Marketing
- Paid Advertising
- Product Marketing
- Public Relations
- Search Engine Optimization
- Social Media Marketing
- Video & Mobile Marketing

CATEGORY 9: Music

SUBCATEGORIES:

- Instruments
- Music Fundamentals
- Music Production
- Music Software
- Music Techniques
- Other Music
- Vocal

CATEGORY 10: Office Productivity

SUBCATEGORIES:

- Apple
- Google
- Microsoft
- Oracle
- Other Office Productivity
- SAP

CATEGORY 11: Personal Development

SUBCATEGORIES:

- Career Development
- Creativity
- Happiness
- Influence
- Leadership
- Memory & Study Skills
- Motivation
- Other Personal Development
- Parenting & Relationships
- Personal Brand Building
- Personal Productivity
- Personal Transformation
- Religion & Spirituality
- Self Esteem & Confidence

```
- Stress Management
-------------------
CATEGORY 12: Photography & Video
SUBCATEGORIES:
- Commercial Photography
- Digital Photography
- Other Photography & Video
- Photography
- Photography Tools
- Portrait Photography
- Video Design
-------------------
CATEGORY 13: Teaching & Academics
SUBCATEGORIES:
- Engineering
- Humanities
- Language Learning
- Math
- Online Education
- Other Teaching & Academics
- Science
- Social Science
- Teacher Training
- Test Prep
-------------------
```

```
#let's check are there unique title of each course
udemy_courses_data.title.nunique()==len(udemy_courses_data)
#We can clearly see that there are some courses with same title
```

```
False
```

Analyze free of cost courses

Let's see that how many courses are offered by udemy free of cost and we'll also check whether those courses have good rating and reviews or not?
Mainly we'll try to analyse all other columns with respect to all free courses offered by Udemy.

```
#first let us see number of courses which are free of cost.
print(f"There are {len(udemy_courses_data)-udemy_courses_data.is_paid.sum()} courses wh
#let's create one new df in which only free courses are there
free_udemy_courses=udemy_courses_data[~udemy_courses_data.is_paid]
print("------------------------------------------------------------")
#we can calculate the percentage of free courses
print(f"There are {round(len(free_udemy_courses)/len(udemy_courses_data)*100,2)}% of th
```

```
There are 21738 courses which are available free of cost
```

```
----------------------------------------------------------
There are 10.36% of the total courses in Udemy which are free of cost.
```

Let us see the top 10 courses which has highest number of subscribers.how many of those courses are free

```
print(f"Out of top 10 courses in the list of highest number of subscribers, {len(udemy_

#We can also see that the highest subscribers course is available zero of cost.
```

```
Out of top 10 courses in the list of highest number of subscribers, 2 courses are free.
```

```
/tmp/ipykernel_39/3964528791.py:3: UserWarning: Boolean Series key will be reindexed to
match DataFrame index.
  print(f"Out of top 10 courses in the list of highest number of subscribers,
{len(udemy_courses_data.sort_values('num_subscribers',ascending=False).head(10)
[udemy_courses_data.price==0])} courses are free.")
```

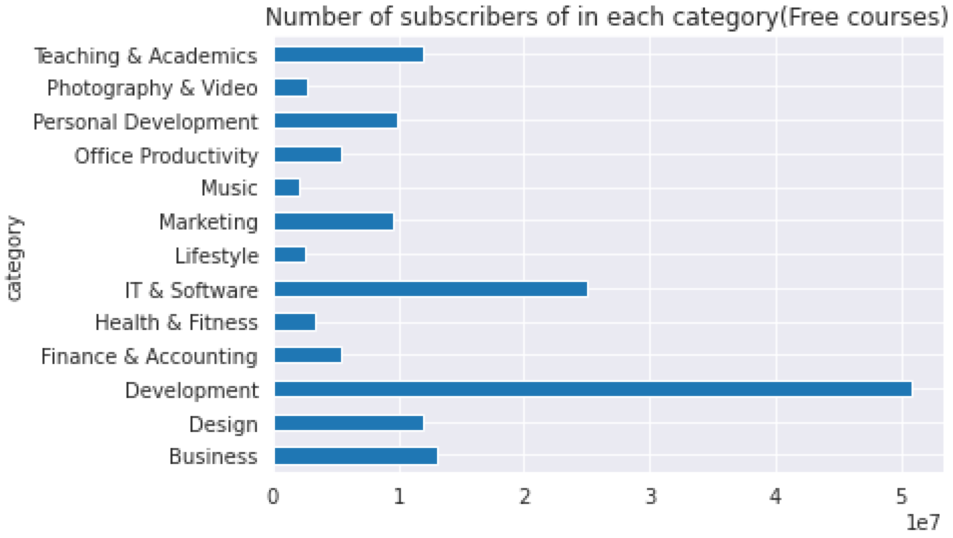
we can see how "category wise the number of subscribers ,avg rating,number of reviews,number of comments,number of lectures and content length in minute" varies by taking average(mean).

```
#first we have group our data with respect to category.We can also
#our new df
category_wise_grouping_of_free_udemy_courses=free_udemy_courses.groupby('category')[['r
category_wise_grouping_of_free_udemy_courses
```

	num_subscribers	num_reviews	num_comments	num_lectures	content_length_min
category					
Business	13049904.0	661873.0	126839.0	37009.0	207977.0
Design	11978213.0	417161.0	87944.0	22948.0	137422.0
Development	50805789.0	2065658.0	364901.0	71962.0	424973.0
Finance & Accounting	5367368.0	268289.0	44041.0	16856.0	95623.0
Health & Fitness	3436233.0	119785.0	28730.0	17243.0	80503.0
IT & Software	25011107.0	1162317.0	198126.0	46596.0	279640.0
Lifestyle	2641592.0	106507.0	28111.0	9446.0	53427.0
Marketing	9641576.0	333573.0	71917.0	21903.0	124047.0
Music	2128164.0	62419.0	14041.0	8128.0	31896.0
Office Productivity	5478065.0	370419.0	72887.0	11127.0	49133.0
Personal Development	9895701.0	508076.0	105217.0	39243.0	208825.0
Photography & Video	2779991.0	99234.0	22090.0	6094.0	29389.0
Teaching & Academics	12012315.0	467604.0	98890.0	48168.0	269633.0

First let's plot number of subscribers of each category and also which category has highest subscribers

```
plt.title('Number of subscribers of in each category(Free courses)')
category_wise_grouping_of_free_udemy_courses.num_subscribers.plot(kind='barh');
```



As expected the most booming field is development so in udemy highest number of subscribers are in Development field.In 2nd IT & Software.WE CAN ALSO SAY THAT MOST LEARNERS ARE IN DEVELOPMENT FIELD.It's quite unexpected that peoples are not interested in music and course related to lifestyle OR it might be the case that peoples are not liking the courses of this categories specifically offered by Udemy.They might be this courses from another platform.
NOTE: this is for courses which are free of cost.

```
#let us check % of the development in free courses
print(f"There over {round(len(udemy_courses_data[(udemy_courses_data.price==0) & (udemy
```

```
There over 15.7% of the courses on development which are free of cost.
```

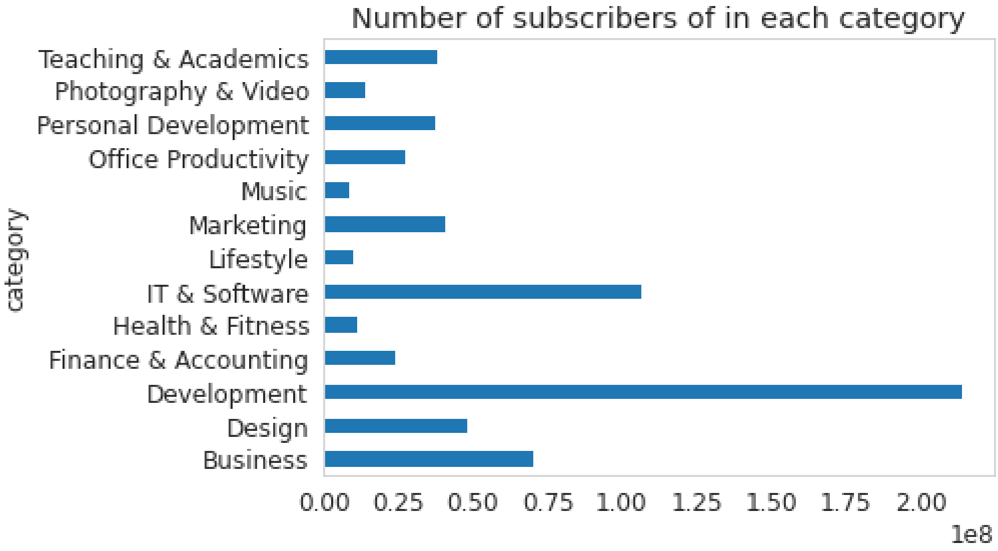
for all courses

```
categoriwise_grouping=udemy_courses_data.groupby('category')[['num_subscribers','num_re
categoriwise_grouping
```

	num_subscribers	num_reviews	num_comments	num_lectures	content_length_min
category					
Business	70012074.0	6962203.0	1209995.0	737666.0	4738894.0
Design	47989137.0	3332380.0	683989.0	615825.0	5206980.0
Development	213749682.0	17200710.0	2882723.0	1961751.0	14914853.0
Finance & Accounting	23822748.0	1832440.0	342988.0	318363.0	2523786.0
Health & Fitness	10967501.0	787346.0	217705.0	297172.0	1797097.0
IT & Software	106766851.0	8560213.0	1408011.0	1085311.0	8248473.0
Lifestyle	10066453.0	996633.0	303878.0	248040.0	1747542.0
Marketing	40803010.0	1962995.0	453389.0	339572.0	2182381.0
Music	8510231.0	691267.0	160306.0	209179.0	1164801.0
Office Productivity	27613503.0	2773102.0	487733.0	306708.0	1898101.0
Personal Development	37214172.0	2922120.0	601759.0	528009.0	3373217.0
Photography & Video	13720894.0	671898.0	165248.0	126747.0	795147.0
Teaching & Academics	38302240.0	2557041.0	494003.0	891098.0	7105449.0

First let's plot number of subscribers of each category and also which category has highest subscribers

```
plt.title('Number of subscribers of in each category')
categoriwise_grouping.num_subscribers.plot(kind='barh');
```

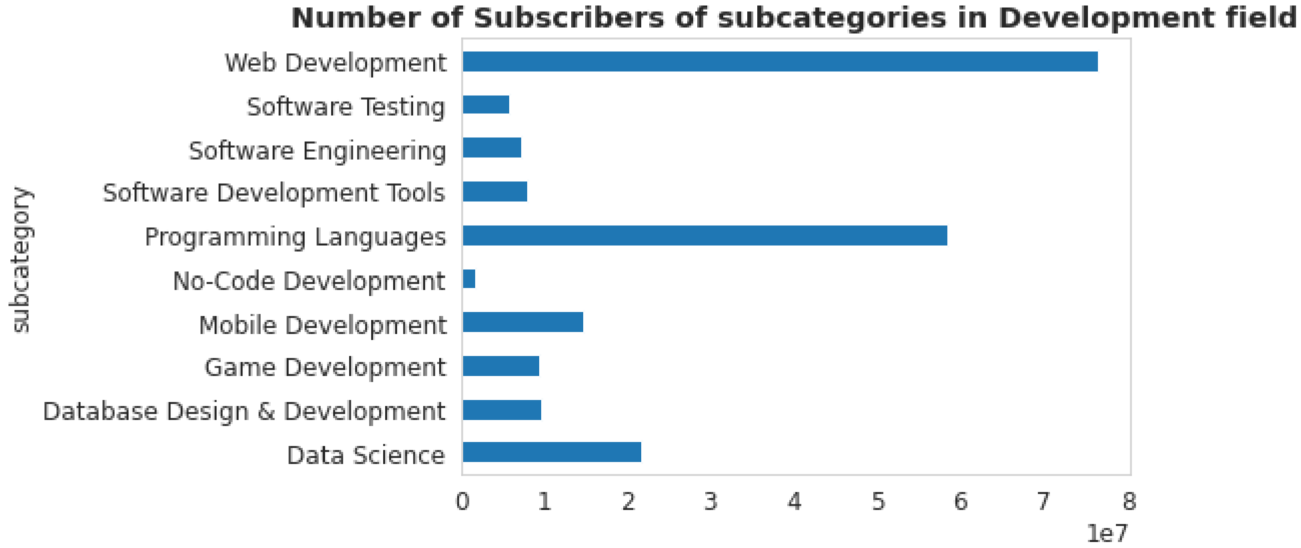


Based on the two above plots the subscribers came from paid courses are big in number as compared to number of subscribers from free courses. You can see in first the range of subscribers from Development category is around 50M, And in overall case of development field number of subscribers are around 200M.

```
#Let's get deep dive in Development field itself (it will be more interesting)
#let's first see the sub categories in deveopment
subcategories_of_Development=udemy_courses_data[udemy_courses_data.category=='Developme
subcategories_of_Development
```

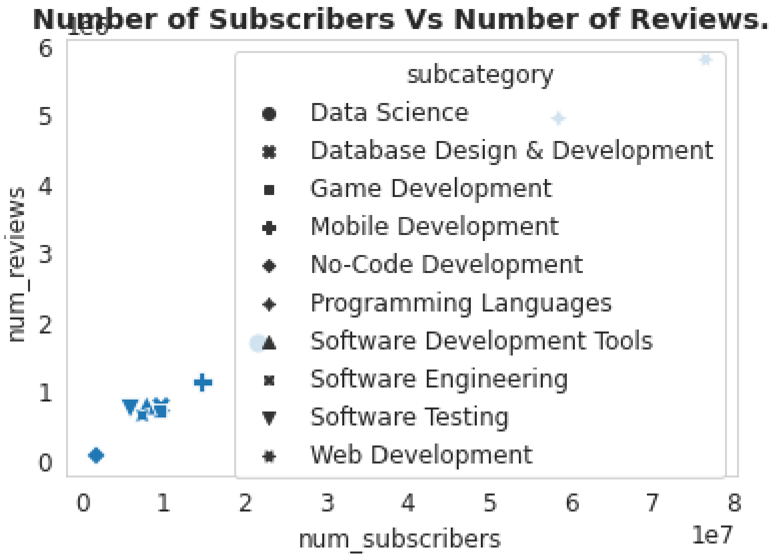
	num_subscribers	avg_rating	num_reviews	num_comments	num_lectures
subcategory					
Data Science	21673262.0	11239.513429	1677710.0	237666.0	179063.0
Database Design & Development	9708331.0	6926.734882	770931.0	123788.0	92105.0
Game Development	9563743.0	8972.320925	700908.0	153469.0	130655.0
Mobile Development	14685346.0	10596.152592	1127305.0	206497.0	193403.0
No-Code Development	1717066.0	2560.545859	53094.0	12562.0	23347.0
Programming Languages	58527936.0	27828.833672	4938064.0	754267.0	448849.0
Software Development Tools	7938135.0	6144.947497	777306.0	125256.0	65595.0
Software Engineering	7391341.0	6146.678994	631181.0	93328.0	78242.0
Software Testing	5917056.0	4288.917478	730646.0	118019.0	56535.0
Web Development	76627466.0	41515.505534	5793565.0	1057871.0	693957.0

```
#let us try to draw with number of subscribers
subcategories_of_Development.num_subscribers.plot(kind='barh')
plt.title('Number of Subscribers of subcategories in Development field',weight='bold');

#We can see that Programming languages and then Data science. Yay! :)
```



In the Development field itself when we compare the subcategories with number of subscribers we found that number of subscribers in web development field around 75M. We are learning data science and our field is in 3rd number in number of subscribers. It might the case that peoples are liking the other platform courses like Jovian.

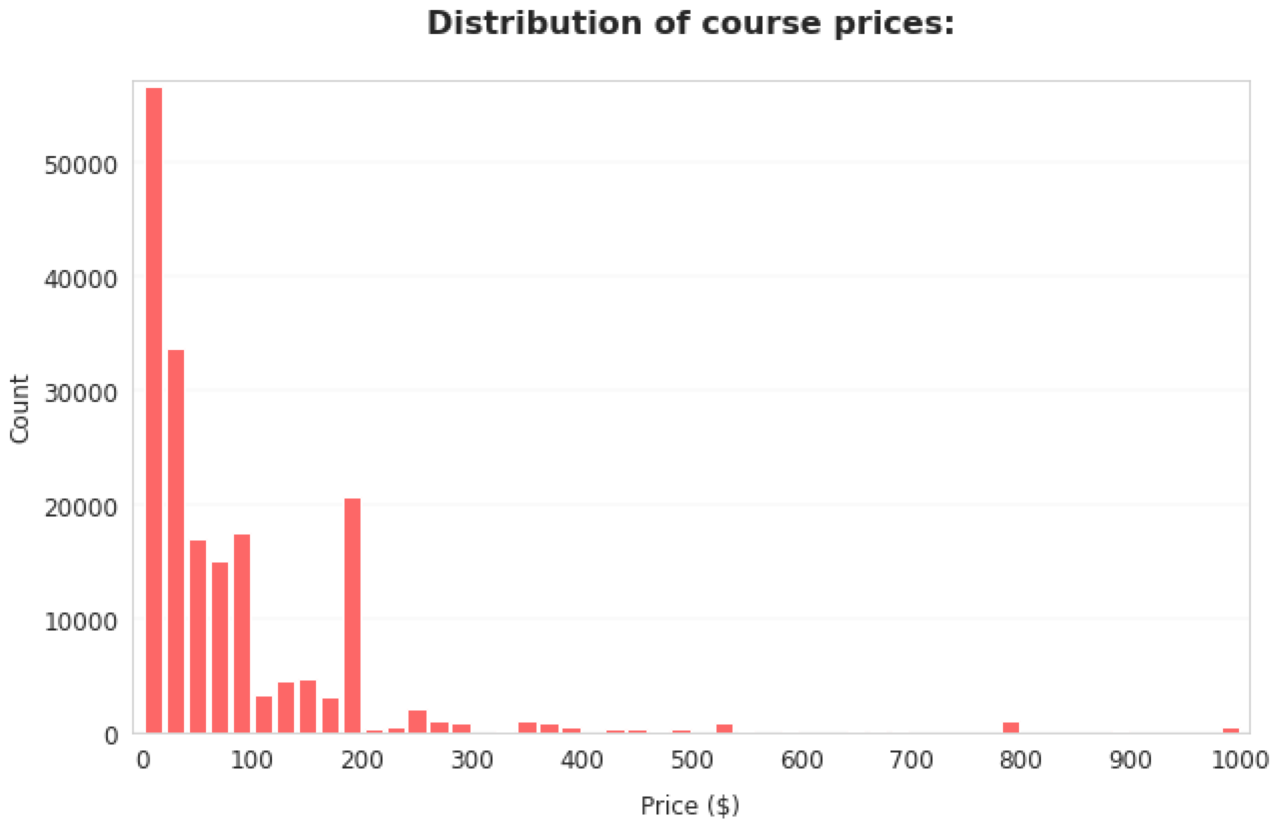
```
#let's compare the number of reviews with number of subscribers
plt.title("Number of Subscribers Vs Number of Reviews.",weight='bold').set_fontsize('14
sns.scatterplot(x='num_subscribers',y='num_reviews',data=subcategories_of_Development,s
```



When we compare the number of subscribers Vs Number of Reviews we can see in above plot Web development field has the highest number of subscribers as well as number of reviews.

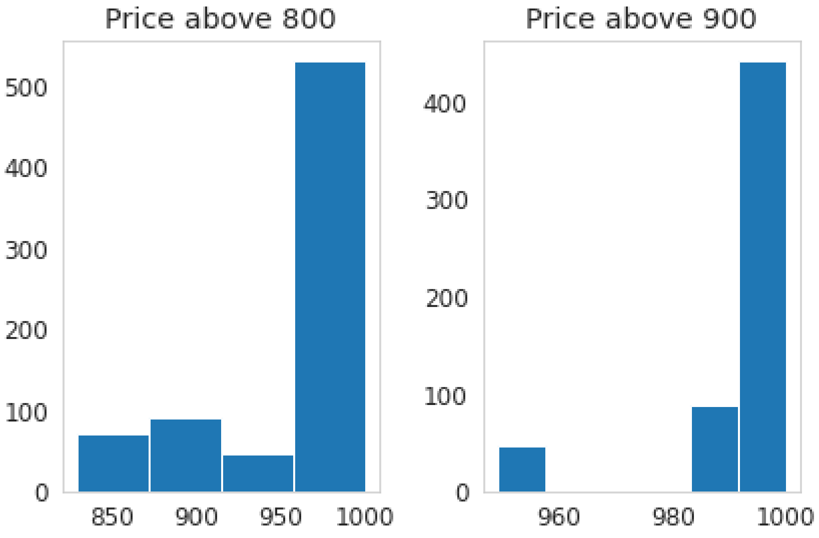
Price distribution

```
#Distribution of price of udemy courses
sns.set_style("whitegrid", {'axes.grid' : False})
plt.rcParams.update({'font.size': 12, 'axes.axisbelow': True})
plt.figure(figsize = (10, 6))
plt.hist(udemy_courses_data[udemy_courses_data['is_paid']==True]['price'],bins=50,range
         color='#fd6767')
plt.grid(axis='y', color ='Grey',
        linestyle ='-.', linewidth = 0.1)
plt.xticks(range(0,1001,100))
plt.margins(0.01)
plt.xlabel("Price ($)", labelpad=10)
plt.ylabel("Count")
plt.title('Distribution of course prices:\n',loc='center',weight='bold', fontdict={'for
plt.show()
```



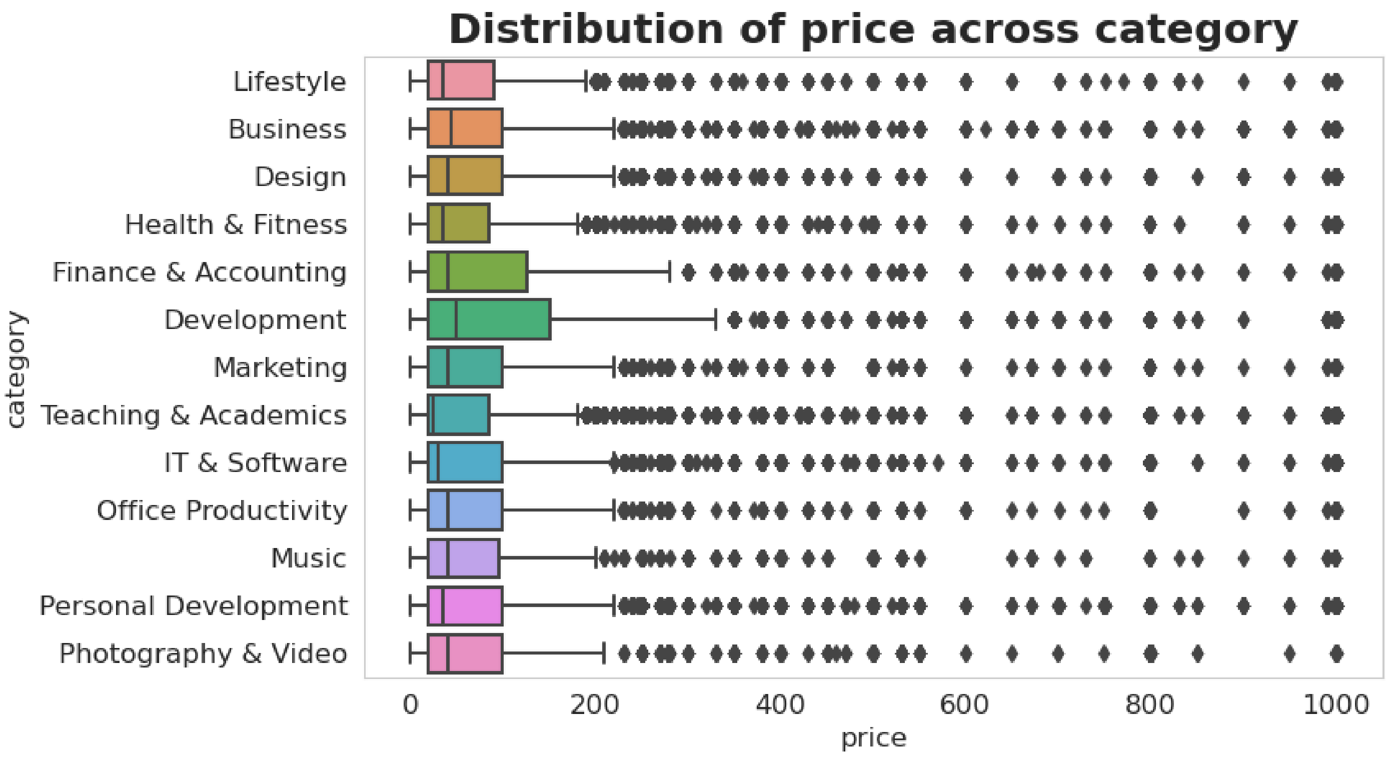
we can see that most of the courses having price between 0 to 100(in this free courses are also available) .There are courses in very less number having price greater than 200.

```
plt.subplot(1,2,1)
plt.title("Price above 800")
plt.hist(udemy_courses_data[udemy_courses_data.price>800].price,bins=4);
plt.subplot(1,2,2)
plt.title("Price above 900")
plt.hist(udemy_courses_data[udemy_courses_data.price>900].price,bins=6);
plt.tight_layout()
```



From the above plot there over 700 courses have price greater than 800. But we can see in 2nd plot ,from them most of the courses having price between 980 to 999.99.

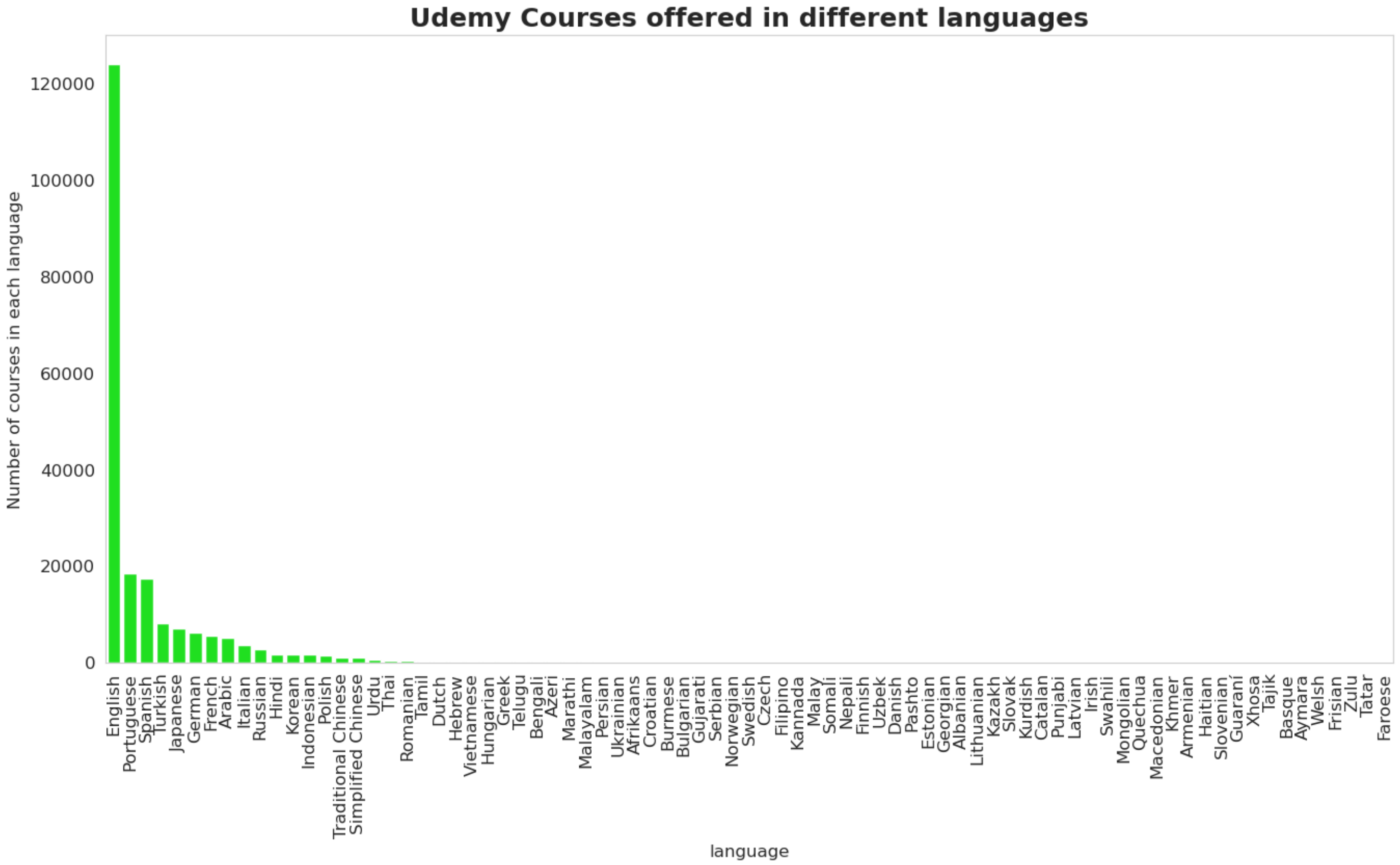
```
# Box plot showing the distribution of price across category
plt.figure(figsize=(8,5), dpi=100)
plt.title("Distribution of price across category",weight='bold').set_fontsize(18)
sns.boxplot(data=udemy_courses_data, x='price', y='category');
```



The above plots show the distribution of the price of Udemy courses as well as the category-wise distribution (boxplot). Most of the courses are priced between 0 and $ 200. Not all the courses are in the paid category. About 21738 courses are offered at no cost whereas 187996 courses have to be bought at a price

Udemy course languages

```
#Let's see the number of courses offered by Udemy in each language.
plt.figure(figsize=(16,8),dpi=100),
sns.countplot(data=udemy_courses_data,x='language',order=udemy_courses_data['language']
plt.ylabel("Number of courses in each language")
plt.title("Udemy Courses offered in different languages",weight='bold').set_fontsize(18
plt.xticks(rotation=90);
```



In English only there are 120000 courses which are in English language itself. And then there is huge gap between the number of courses in English and all other languages.

```
def calculate_percentage(lang):
    return round(len(udemy_courses_data[udemy_courses_data.language==lang])*100/len(ude
```

```
print(f'English({calculate_percentage("English")}%),Portuguese({calculate_percentage("P
```

```
English(59.08%),Portuguese(8.81%), and Spanish(8.31%) are the top three languages on
Udemy
```

The top 15 languages in which the courses are offered are visualized in the pie chart shown below.

But there are so many languages so pie plot will not be comfortable so we will use plotly.

To use it first we have to install it so
let's install the plotly as px

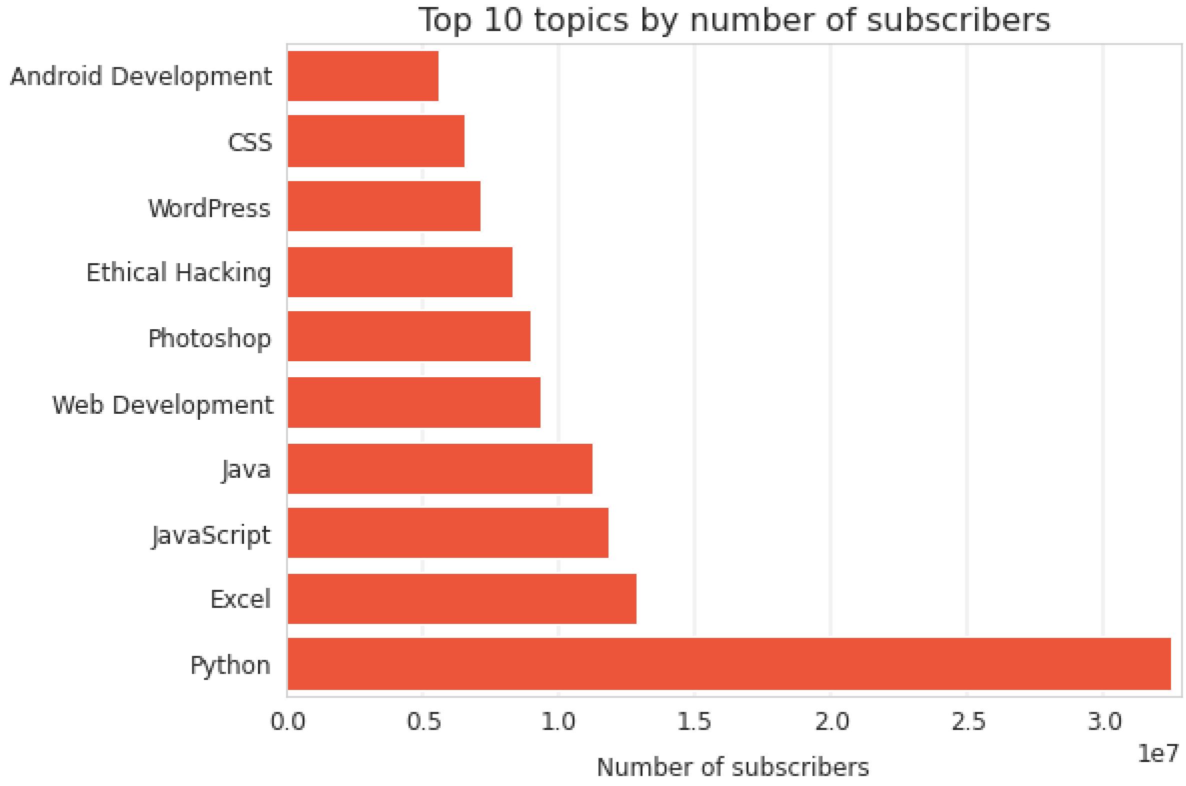
```
we can also use plotly
!pip install plotly --upgrade --quiet
import plotly
import plotly.express as px
fig=px.pie(share_of_languages_top_15,values=share_of_languages_top_15.values,names=shar
fig.update_traces(textposition="inside",texttemplate='%{label}<br>%{value}%',rotation=1
fig.update_layout(title_text='Top 15 languages in which courses are offered', title_x=0
fig.show()
```

Based on the course dataset, there are 79 languages in which Udemy courses are offered. English (59%), Portugues (8.8%), and Spanish (8.3%) are the top three languages on Udemy.

```
#let's see the top 10 topics by number of subscribers
topic_top_10=udemy_courses_data.groupby(['topic']).sum(numeric_only=True).sort_values(b
print("Top 10 topic by number of subscribers: \n")
print(topic_top_10['num_subscribers'])
```

```
Top 10 topic by number of subscribers:


topic
Python                 32516280.0
Excel                  12822452.0
JavaScript             11801744.0
Java                   11203723.0
Web Development         9293697.0
Photoshop               8976024.0
Ethical Hacking         8280273.0
WordPress               7080627.0
CSS                     6482557.0
Android Development     5567763.0
Name: num_subscribers, dtype: float64
```

```
fig,ax=plt.subplots(figsize=(8,6))
ax.barh(topic_top_10.index,topic_top_10['num_subscribers'].values,color='#ed553b')

plt.title('Top 10 topics by number of subscribers',fontsize=16)
plt.grid(axis='x',color='Grey',linestyle='-',linewidth=0.2)
plt.xlabel('Number of subscribers',labelpad=10)
plt.margins(0.01)
```



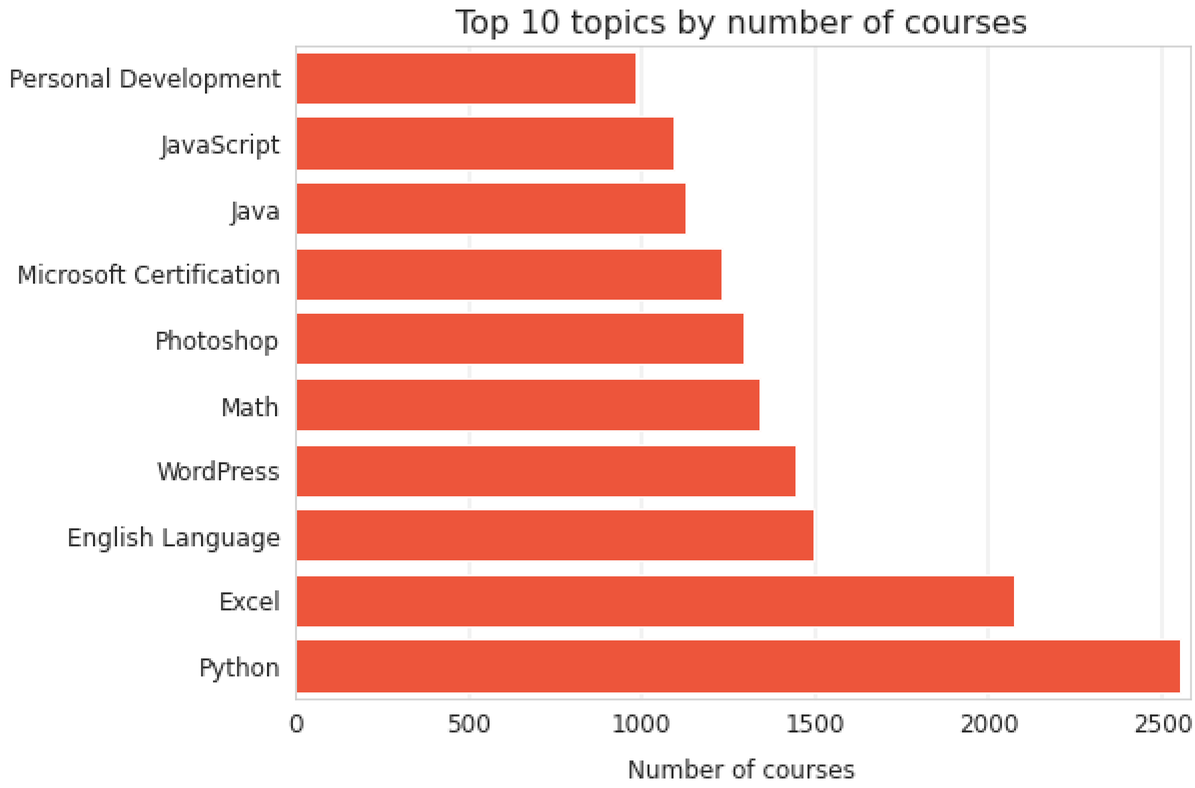
Above plot shows that the number of subscribers for python are much as compared to others.
Note: There are courses which has same topic names and subcategory name.

```
#Top 10 topics by number of courses
topic_df_top_10 = udemy_courses_data.groupby(['topic'],).size().sort_values(ascending=F
print('Top 10 topics by number of courses: \n')
print(topic_df_top_10)
```

```
Top 10 topics by number of courses:


topic
Python                    2553
Excel                     2072
English Language          1495
WordPress                 1442
Math                      1341
Photoshop                 1294
Microsoft Certification   1232
Java                      1128
JavaScript                1092
```

```
Personal Development       978
dtype: int64
```

```
fig,ax=plt.subplots(figsize=(8,6))
ax.barh(topic_df_top_10.index,topic_df_top_10.values,color='#ed553b')

plt.title('Top 10 topics by number of courses',fontsize=16)
plt.grid(axis='x',color='Grey',linestyle='-',linewidth=0.2)
plt.xlabel('Number of courses',labelpad=10)
plt.margins(0.01)
```



Based on the above plot the courses are offered from udemy is on python.Then there is much more difference between number of courses in python which topmost and on Excel which is 2nd topmost.There is like linear decrement from Excel courses onward till Android Development.

Top 5 courses with highest number of reviews,subscribers,comments,content length,etc by using function.

```
#let's first select the columns which we need and then we can define one function which
recq_column=['title','instructor_name','num_lectures','avg_rating','price','num_reviews

def top_courses_5(df,col):
    top5=df.nlargest(5,col).reset_index(drop=True)
    return top5[recq_column]
```

```
print(f"Top 5 courses with highest number of reviews are")
top_courses_5(udemy_courses_data,'num_reviews')
```

```
Top 5 courses with highest number of reviews are
```

	title	instructor_name	num_lectures	avg_rating	price	num_reviews	num_subscribers	num_comments	c
0	2022 Complete Python Bootcamp From Zero to Her...	Jose Portilla	255.0	4.611872	199.99	436457.0	1612862.0	39040.0	
1	Microsoft Excel - Excel from Beginner to Advanced	Kyle Pew	269.0	4.689838	149.99	332598.0	1108811.0	36101.0	
2	The Web Developer Bootcamp 2022	Colt Steele	648.0	4.708899	199.99	246624.0	823805.0	31001.0	
3	The Complete 2022 Web Development Bootcamp	Dr. Angela Yu	531.0	4.698089	199.99	228108.0	771176.0	27723.0	
4	Angular - The Complete Guide (2022 Edition)	Maximilian Schwarzmüller	698.0	4.646908	189.99	172991.0	626304.0	24886.0	

```
print(f"Top 5 courses with highest number of subscribers are")
top_courses_5(udemy_courses_data,'num_subscribers')
```

```
Top 5 courses with highest number of subscribers are
```

	title	instructor_name	num_lectures	avg_rating	price	num_reviews	num_subscribers	num_comments	c
0	Java Tutorial for Complete Beginners	John Purcell	74.0	4.444383	0.00	96582.0	1752364.0	14919.0	
1	2022 Complete Python Bootcamp From Zero to Her...	Jose Portilla	255.0	4.611872	199.99	436457.0	1612862.0	39040.0	
2	Microsoft Excel - Excel from Beginner to Advanced	Kyle Pew	269.0	4.689838	149.99	332598.0	1108811.0	36101.0	
3	Automate the Boring Stuff with Python Programming	Al Sweigart	51.0	4.676586	49.99	102876.0	1056369.0	13544.0	
4	Machine Learning A-Z™: Hands-On Python & R In ...	Kirill Eremenko	340.0	4.569116	199.99	162432.0	896340.0	22567.0	

```
print(f"Top 5 courses with highest number of comments are")
top_courses_5(udemy_courses_data,'num_comments')
```

```
Top 5 courses with highest number of comments are
```

	title	instructor_name	num_lectures	avg_rating	price	num_reviews	num_subscribers	num_comments	cc
0	2022 Complete Python Bootcamp From Zero to Her...	Jose Portilla	255.0	4.611872	199.99	436457.0	1612862.0	39040.0	
1	Microsoft Excel - Excel from Beginner to Advanced	Kyle Pew	269.0	4.689838	149.99	332598.0	1108811.0	36101.0	
2	The Web Developer Bootcamp 2022	Colt Steele	648.0	4.708899	199.99	246624.0	823805.0	31001.0	
3	The Complete 2022 Web Development Bootcamp	Dr. Angela Yu	531.0	4.698089	199.99	228108.0	771176.0	27723.0	
4	The Complete Digital Marketing Course - 12 Cou...	Rob Percival	411.0	4.510407	199.99	154985.0	706339.0	27540.0	

```python
print(f"Top 5 courses with highest content length in minute are")
top_courses_5(udemy_courses_data,'content_length_min')
```

```
Top 5 courses with highest content length in minute are
```

	title	instructor_name	num_lectures	avg_rating	price	num_reviews	num_subscribers	num_comments
0	Chemistry for IIT JEE Main & Advanced, NEET, A...	Aman Saurav	354.0	4.350000	7.00	76.0	463.0	5.0
1	Crush Your 2019 New Year's Resolution and Lear...	Mammoth Interactive	342.0	4.750000	199.99	8.0	333.0	3.0
2	NET ENGLISH COMPLETE COURSE	Kalyani Vallath	417.0	4.851852	7.00	623.0	2397.0	78.0
3	Comprehensive Human Psychology Course	Bilal Semih Bozdemir	740.0	3.875000	99.99	4.0	536.0	1.0
4	Kapsamlı Psikoloji Kursu	Bilal Semih Bozdemir	792.0	3.727273	999.99	328.0	2384.0	54.0

```python
print(f"Top 5 courses with most expensive are")
top_courses_5(udemy_courses_data,'price')
```

```
Top 5 courses with most expensive are
```

	title	instructor_name	num_lectures	avg_rating	price	num_reviews	num_subscribers	num_comments
0	Oracle Veritabanı Programlama : SQL, PL/SQL, O...	Cihan Özhan	280.0	4.842857	999.99	570.0	32834.0	93.0
1	React Native ile iOS ve Android Uygulama Geliş...	Yasin Ugurlu	110.0	3.900000	999.99	838.0	2940.0	141.0
2	Sıfırdan Wordpress Tema Yapımı	Ali Çınaroğlu	27.0	4.100000	999.99	206.0	1015.0	50.0
3	Go Programlama Dili	Cihan Özhan	97.0	4.033333	999.99	625.0	37154.0	79.0
4	SQL Server Veritabanı Programlama: Temel, Orta...	Cihan Özhan	258.0	4.492064	999.99	1348.0	49938.0	151.0

```
#we can see here all 5 courses have same price which is $999.99 so we can
#check that how many courses have price equal to $999.99.
print(f'There are total {len(udemy_courses_data[udemy_courses_data.price==999.99])} num
```

```
There are total 129 number of courses having price equal to $999.99.
```

```
print(f"Top 5 courses with most average rating are")
top_courses_5(udemy_courses_data,'avg_rating')
```

```
Top 5 courses with most average rating are
```

	title	instructor_name	num_lectures	avg_rating	price	num_reviews	num_subscribers	num_comments	co
0	From Startup to Millions Before 30: Part 2	TeleTime Productions	7.0	5.0	34.99	1.0	72.0	0.0	
1	Two-layered online form validation with jQuery...	Sebastian Sulinski	18.0	5.0	19.99	6.0	78.0	2.0	
2	Win Them Over with Web Video Part 2	Kamala Appel	54.0	5.0	34.99	3.0	53.0	3.0	
3	American Accent Training for IT Professionals	Susan Ryan	117.0	5.0	49.99	340.0	3456.0	53.0	
4	Personal SEO : Become a Creative Brand Advocate	Philip Campbell	67.0	5.0	19.99	2.0	339.0	1.0	

```
print(f"Top 5 courses with highest number of lectures are")
top_courses_5(udemy_courses_data,'num_lectures')
```

```
Top 5 courses with highest number of lectures are
```

	title	instructor_name	num_lectures	avg_rating	price	num_reviews	num_subscribers	num_commen
0	Certified Information Systems Security Profess...	Integrity Training	1095.0	4.350962	199.99	328.0	2728.0	61
1	React - The Complete Guide (incl Hooks, React ...	Academind by Maximilian Schwarzmüller	987.0	4.648803	189.99	163324.0	635960.0	18290
2	(130+Saat)Komple Uygulamalı Web Geliştirme Eği...	Can Boz	961.0	4.532787	169.99	350.0	2008.0	55
3	Microsoft Office 2013 Training Tutorial	TeachUcomp, Inc.	892.0	4.000000	24.99	35.0	366.0	10
4	100+ Saatlik Komple Frontend Eğitimi | Web Tas...	Can Boz	839.0	4.591464	169.99	633.0	10021.0	65

Asking Questions and Answers

1]What are top 10 courses with best ratios of time passed since publishing and the number of subscribers?(Which are top 10 fastest growing courses?)

To answer first question we need to calculate the number of days since the published date of the course and date of Oct 10 2022 and then we have to calculate the ratio of number of subscribers and time passed since publishing.The courses having higher magnitude of ratio, those are the top courses.

```
#let's first import datetime to change datatype into datetime
import datetime
udemy_courses_data['days_since_publishing']=datetime.date(2022,10,10)-udemy_courses_dat
#let's have a quick look at the results
print(udemy_courses_data['days_since_publishing'][:10])
print('\n')
#let's convert this into integer type from time type to perform more operations on that
udemy_courses_data['days_since_publishing']=udemy_courses_data['days_since_publishing']
print(udemy_courses_data['days_since_publishing'][:10])
```

```
0    4449 days
1    4654 days
2    4380 days
3    4130 days
4    4127 days
5    4105 days
6    4562 days
7    4112 days
```

```
8   4114 days
9   3751 days
Name: days_since_publishing, dtype: timedelta64[ns]
```

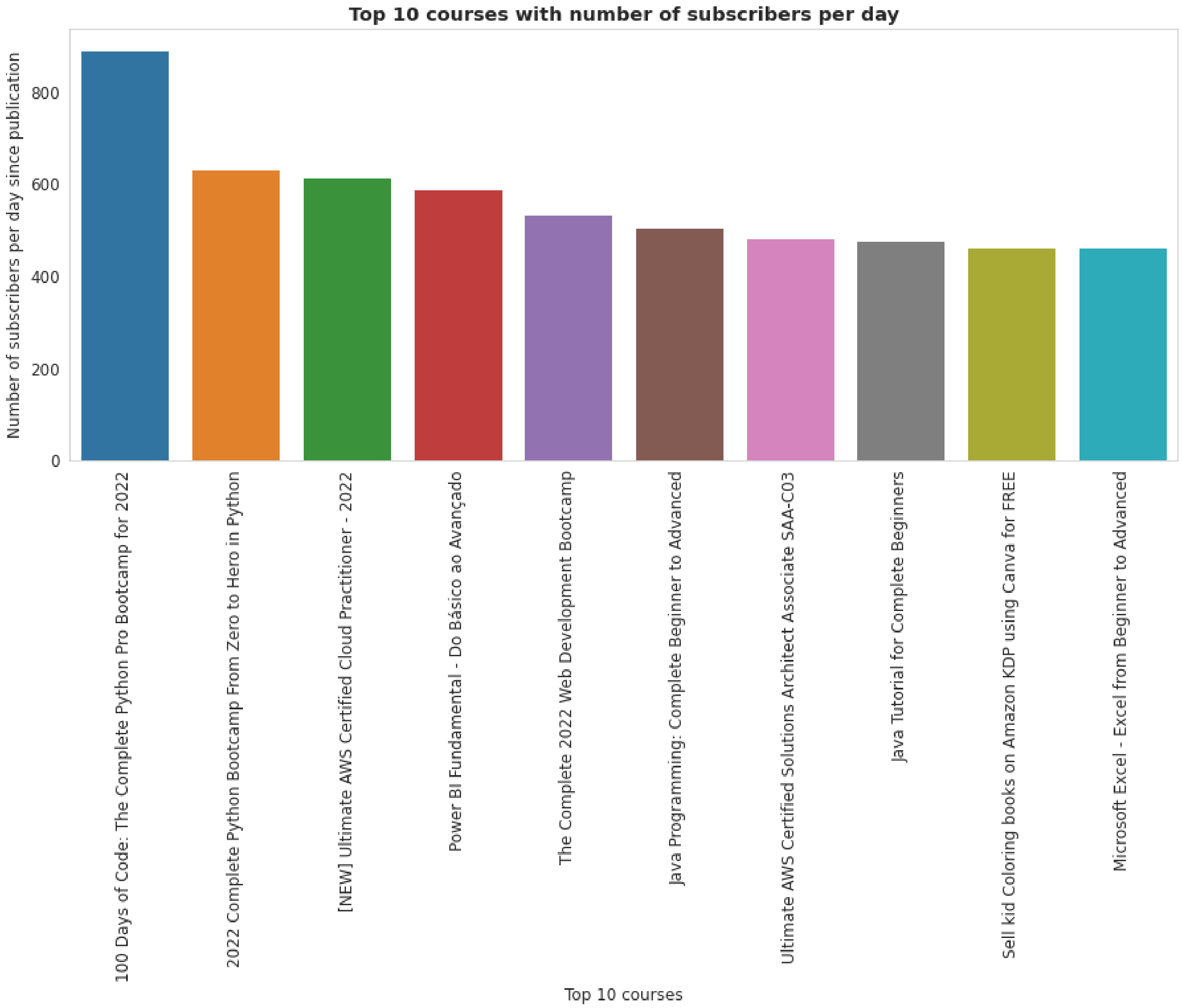
```
0    4449
1    4654
2    4380
3    4130
4    4127
5    4105
6    4562
7    4112
8    4114
9    3751
Name: days_since_publishing, dtype: int64
```

```
#let's quick look through our data
udemy_courses_data['days_since_publishing'].describe()
```

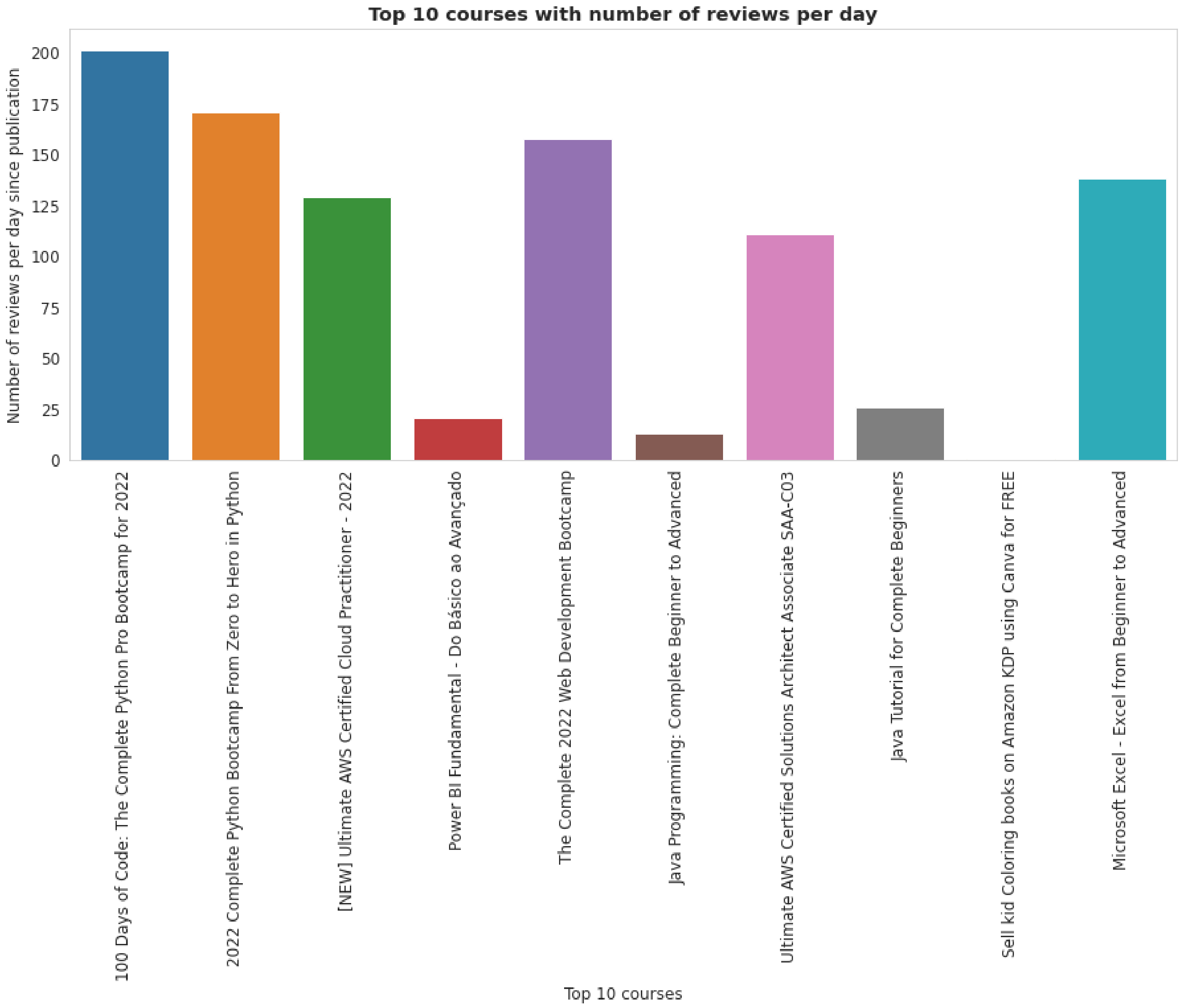
```
count    209734.000000
mean        976.922955
std         761.644872
min           5.000000
25%         406.000000
50%         780.000000
75%        1402.000000
max        4654.000000
Name: days_since_publishing, dtype: float64
```

```
#let's calculate the ratio and store the data in new column.
udemy_courses_data['subscribers_per_day']=udemy_courses_data['num_subscribers']/udemy_c
udemy_courses_data['reviews_per_day']=udemy_courses_data['num_reviews']/udemy_courses_d
udemy_courses_data['comments_per_day']=udemy_courses_data['num_comments']/udemy_courses
udemy_courses_data['lectures_per_day']=udemy_courses_data['num_lectures']/udemy_courses
#let's sort those ratios in descending order and show some perticular column which are
Top_10_fastest_growing_courses=udemy_courses_data[['title','instructor_name','subscribe
```

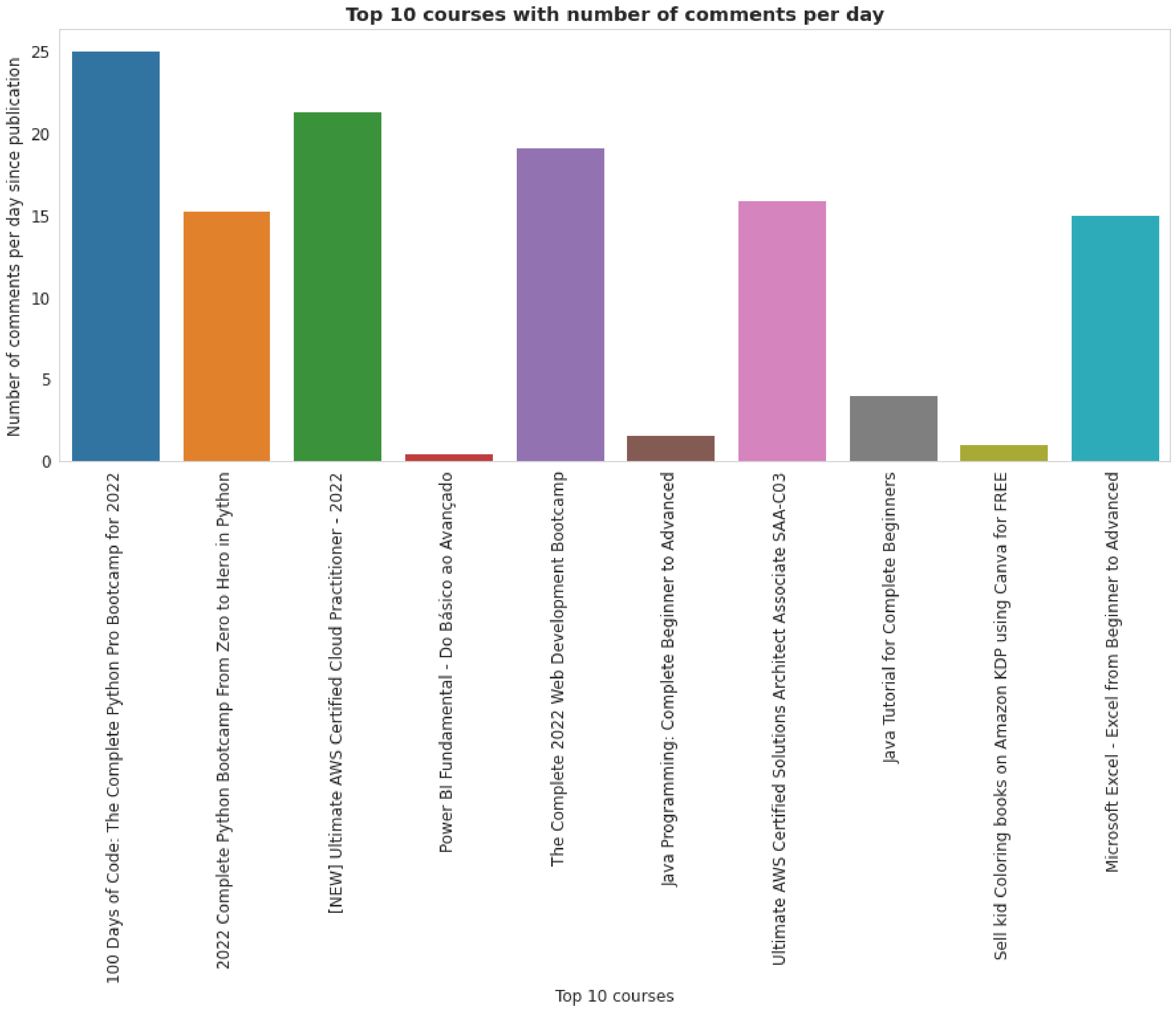
```
#let's also draw the plot by using seaborn
fig,ax=plt.subplots(figsize=(15,6))
plot=sns.barplot(data=Top_10_fastest_growing_courses,x='title',y='subscribers_per_day')
plot.tick_params(axis='x',rotation=90)
plt.title("Top 10 courses with number of subscribers per day",weight='bold')
plot.set_xlabel("Top 10 courses")
plot.set_ylabel("Number of subscribers per day since publication");
```



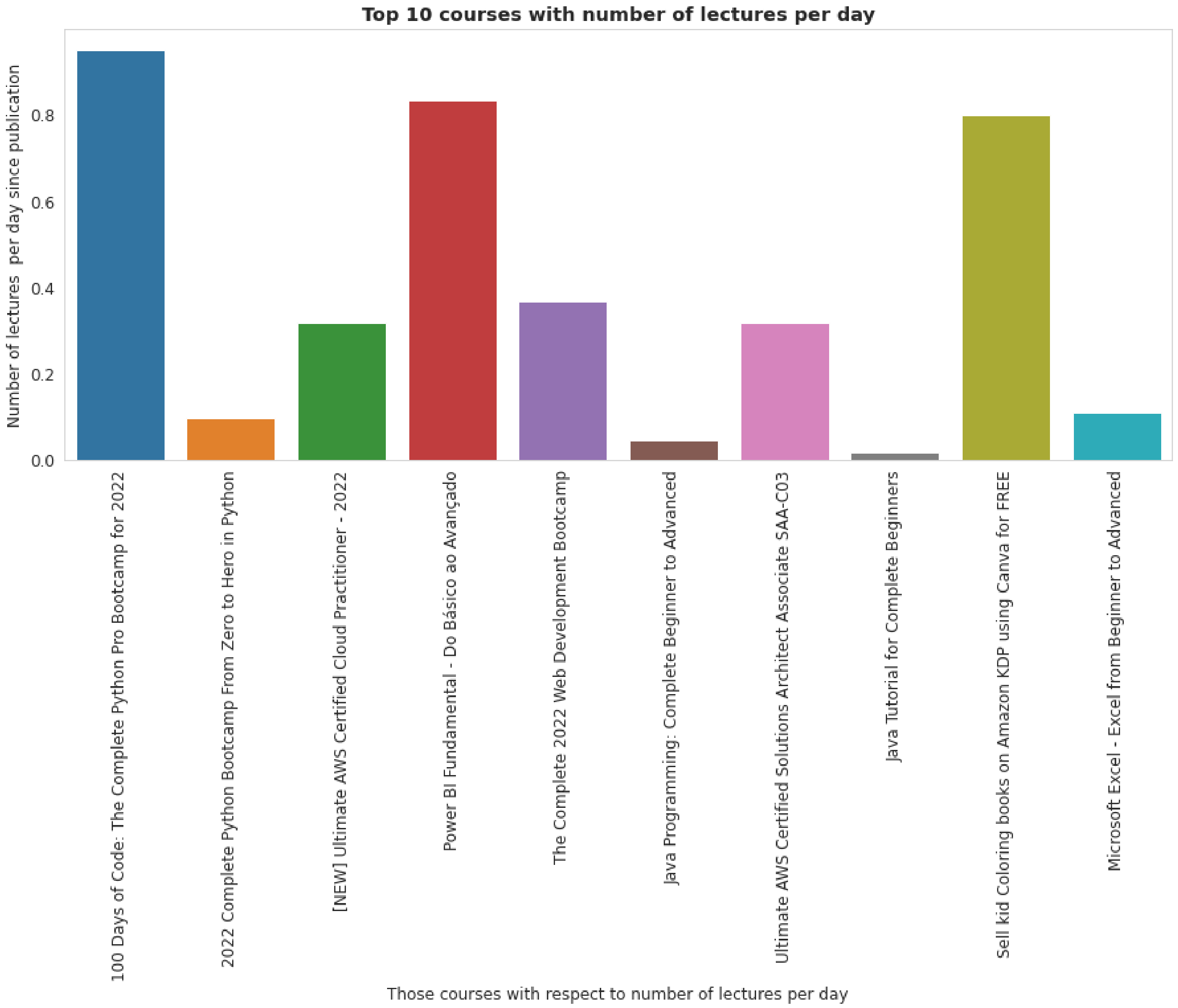
```
#let's also check this courses have number of reviews as compare to subscribers
fig,ax=plt.subplots(figsize=(15,6))
plot=sns.barplot(data=Top_10_fastest_growing_courses,x='title',y='reviews_per_day')
plot.tick_params(axis='x',rotation=90)
plt.title("Top 10 courses with number of reviews per day",weight='bold')
plot.set_xlabel("Top 10 courses")
plot.set_ylabel("Number of reviews per day since publication");


#when we compared those courses with number of reviews 5 courses out of 10 have
#Very less number of reviews as compared to number of subscribers
```



```
#let's also check with number of comments
fig,ax=plt.subplots(figsize=(15,6))
plot=sns.barplot(data=Top_10_fastest_growing_courses,x='title',y='comments_per_day')
plot.tick_params(axis='x',rotation=90)
plt.title("Top 10 courses with number of comments per day",weight='bold')
plot.set_xlabel("Top 10 courses")
plot.set_ylabel("Number of comments per day since publication");
```



```
#let's also check with number of lectures
fig,ax=plt.subplots(figsize=(15,6))
plot=sns.barplot(data=Top_10_fastest_growing_courses,x='title',y='lectures_per_day')
plot.tick_params(axis='x',rotation=90)
plt.title("Top 10 courses with number of lectures per day",weight='bold')
plot.set_xlabel("Those courses with respect to number of lectures per day")
plot.set_ylabel("Number of lectures  per day since publication");
```



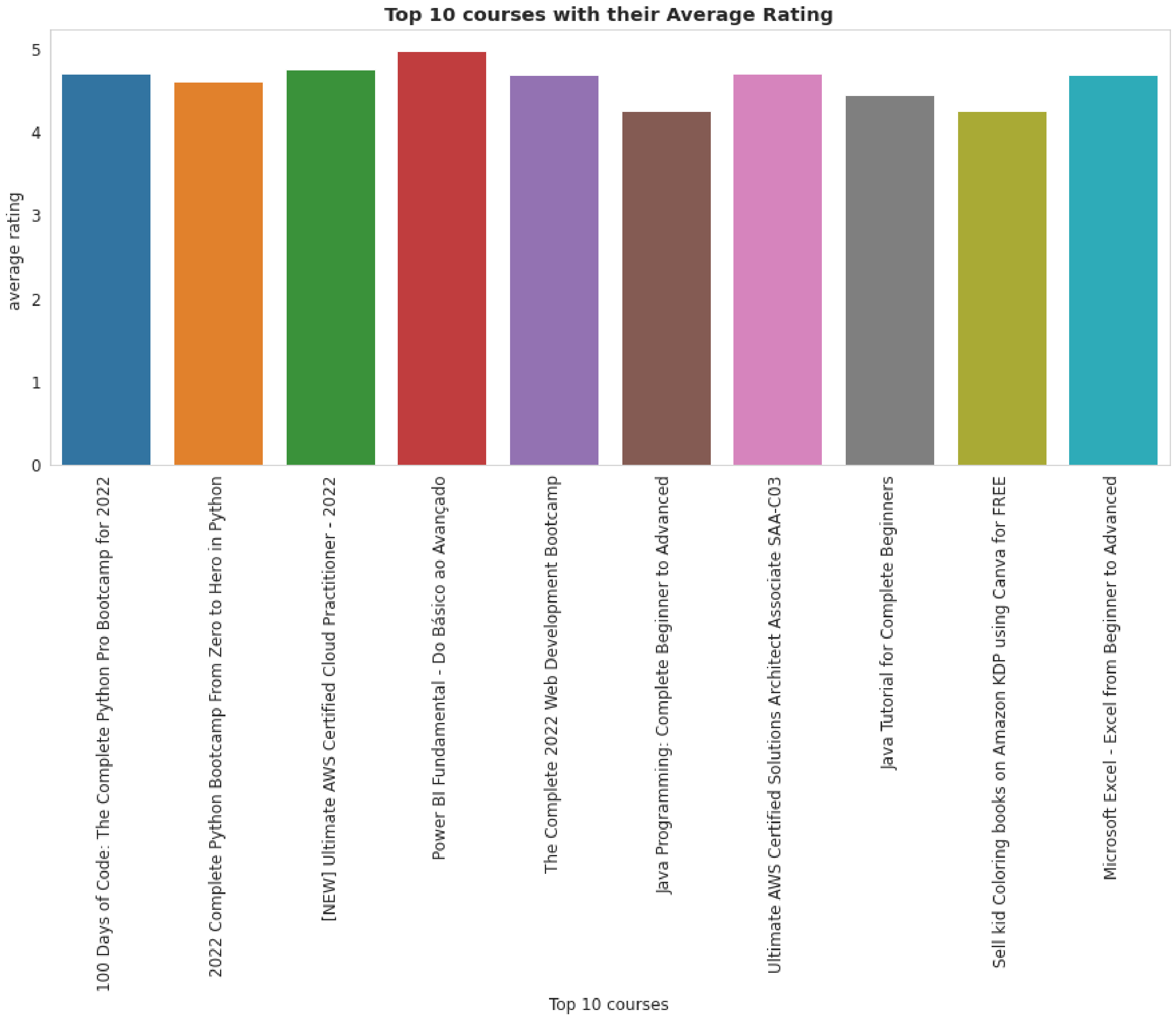
From the above data we can say that 1st course(Which was instructed by Dr. Angela Yu) which has highest number of subscribers per day have great number of reviews,comments in per day.
So the course '100 Days of Code: The Complete Python Pro Bootcamp for 2022' is most growing course.

On the other hand courses 'Professional in advance PHP Language - Practice Test 2022','Become A Certified JavaScript Developer Practice Tests 2022' has not satisfying number umber of reviews as compared to number of subscribers.

The courses named 'Java Tutorial for Complete Beginners','Power BI Fundamental - Do Básico ao Avançado' are free of cost available on Udemy.

let's also the avg rating of this courses

```
fig,ax=plt.subplots(figsize=(15,6))
plot=sns.barplot(data=Top_10_fastest_growing_courses,x='title',y='avg_rating')
plot.tick_params(axis='x',rotation=90)
plt.title("Top 10 courses with their Average Rating",weight='bold')
plot.set_xlabel("Top 10 courses")
plot.set_ylabel("average rating ");
```



Did you notice that the 7th column is always showing zero value except on number of subscribers.

```
udemy_courses_data[udemy_courses_data.title=='Professional in advance PHP Language - Pr

#What is this?
#There is which has zero number of lectures still there are number of
#subscribers so this row needs to be drop.

#WAIT...WAIT....
#Are there another courses which has zero number of lectures
#let's check
```

	id	title	is_paid	price	headline	num_subscribers	avg_rating	num_reviews	num_comment
209718	4912856.0	Professional in advance PHP Language - Practic...	True	19.99	The Complete PHP Developer Course Exam	3001.0	0.0	0.0	0.

1 rows × 25 columns

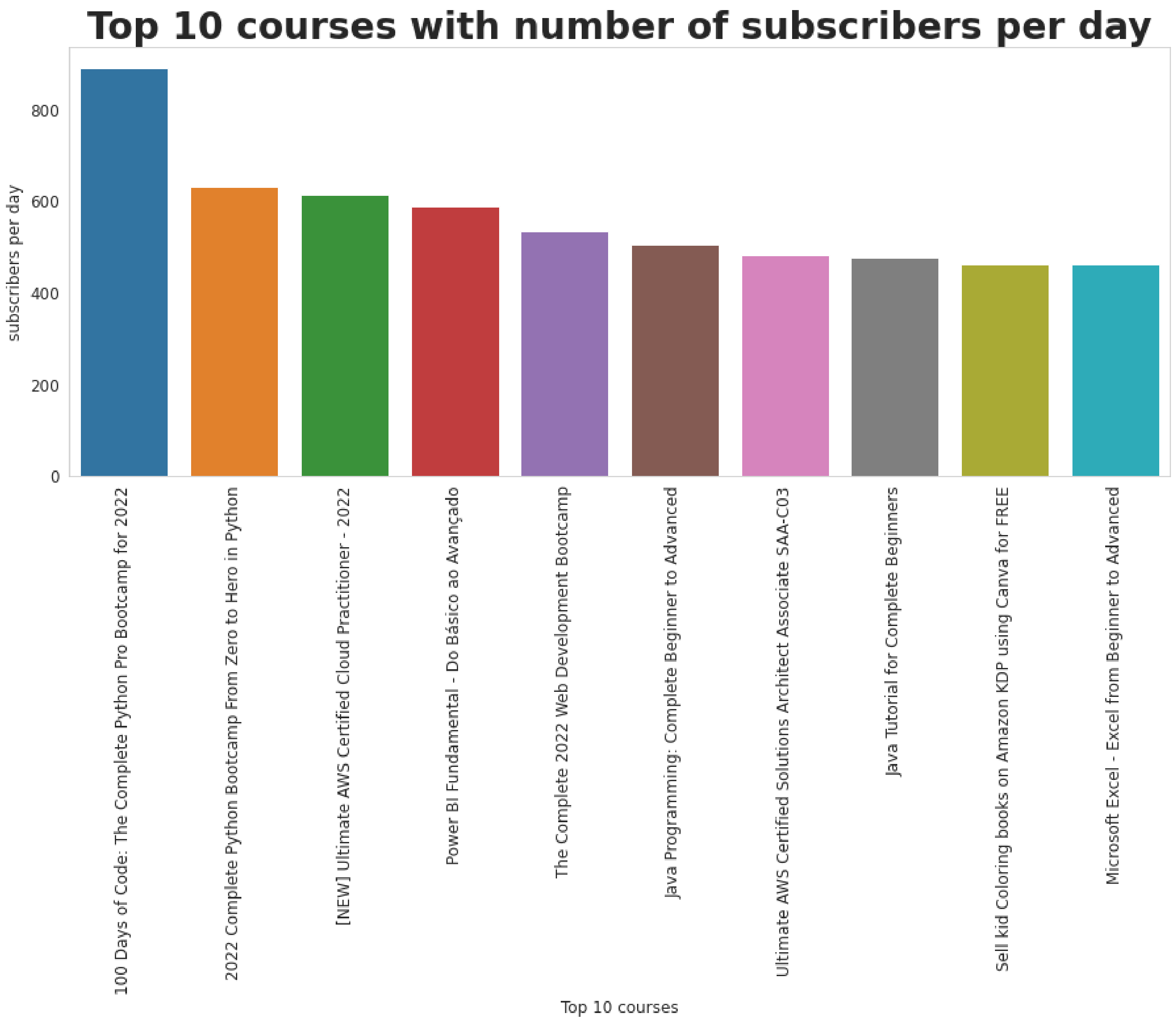
```
#so we will check those with zero number of lectures's.
Need_to_drop_columns=udemy_courses_data[(udemy_courses_data.num_lectures==0)]
print(f'There are {len(Need_to_drop_columns)} rows in our Udemy courses data which can
```

```
There are 9337 rows in our Udemy courses data which can be clean.
```

```
#let's first drop those columns from this data.
udemy_courses_data.drop(index=[209718,209721],inplace=True)
```

```
#let's again define our df
Top_10_fastest_growing_courses=udemy_courses_data[['title','instructor_name','subscribe
```

```
fig,ax=plt.subplots(figsize=(15,6))
plot=sns.barplot(data=Top_10_fastest_growing_courses,x='title',y='subscribers_per_day')
plot.tick_params(axis='x',rotation=90)
plt.title("Top 10 courses with number of subscribers per day",weight='bold').set_fontsi
plot.set_xlabel("Top 10 courses")
plot.set_ylabel("subscribers per day");


#So :) our answer is ready..........:)
```



In the above plot we can see the top 10 courses .

2]Top 10 most popular instructors?

```
#our answer is in the first question itself
Top_10_fastest_growing_courses[['title','instructor_name']].reset_index()
```

	index	title	instructor_name
0	84101	100 Days of Code: The Complete Python Pro Boot...	Dr. Angela Yu
1	10724	2022 Complete Python Bootcamp From Zero to Her...	Jose Portilla
2	98405	[NEW] Ultimate AWS Certified Cloud Practitione...	Stephane Maarek | AWS Certified Cloud Practiti...
3	208170	Power BI Fundamental - Do Básico ao Avançado	Stefano Larmelina
4	39113	The Complete 2022 Web Development Bootcamp	Dr. Angela Yu
5	81128	Java Programming: Complete Beginner to Advanced	CodeIn Academy
6	60269	Ultimate AWS Certified Solutions Architect Ass...	Stephane Maarek | AWS Certified Cloud Practiti...
7	396	Java Tutorial for Complete Beginners	John Purcell
8	209472	Sell kid Coloring books on Amazon KDP using Ca...	Passive Income Gen Z

	index	title	instructor_name
9	16288	Microsoft Excel - Excel from Beginner to Advanced	Kyle Pew

In the first question we analyze about the top 10 courses by number of subscribers, and their comments,their lectures,their reviews and all.Similarly we answered the 2nd question.

3]The top 10 instructors who earns highest from their courses?

```
# here first we have to do gross sale analysis.let's calulate total earning of udemy
print(f"Total gross sales of udemy: {round((udemy_courses_data['price']*udemy_courses_c
```

```
Total gross sales of udemy: 59.93 billion US Dollar
```

```
#let's first create a column of earning by each course
udemy_courses_data['earning']=udemy_courses_data.price*udemy_courses_data.num_subscribe
instructor_earning=udemy_courses_data.groupby('instructor_name')['earning'].sum().rounc
instructor_earning=instructor_earning.sort_values(ascending=False)
instructor_earning=instructor_earning.apply(lambda x: "{:,}".format(x))
print('Top 10 instructors with their earnings in $:')
instructor_earning.head(10)
```

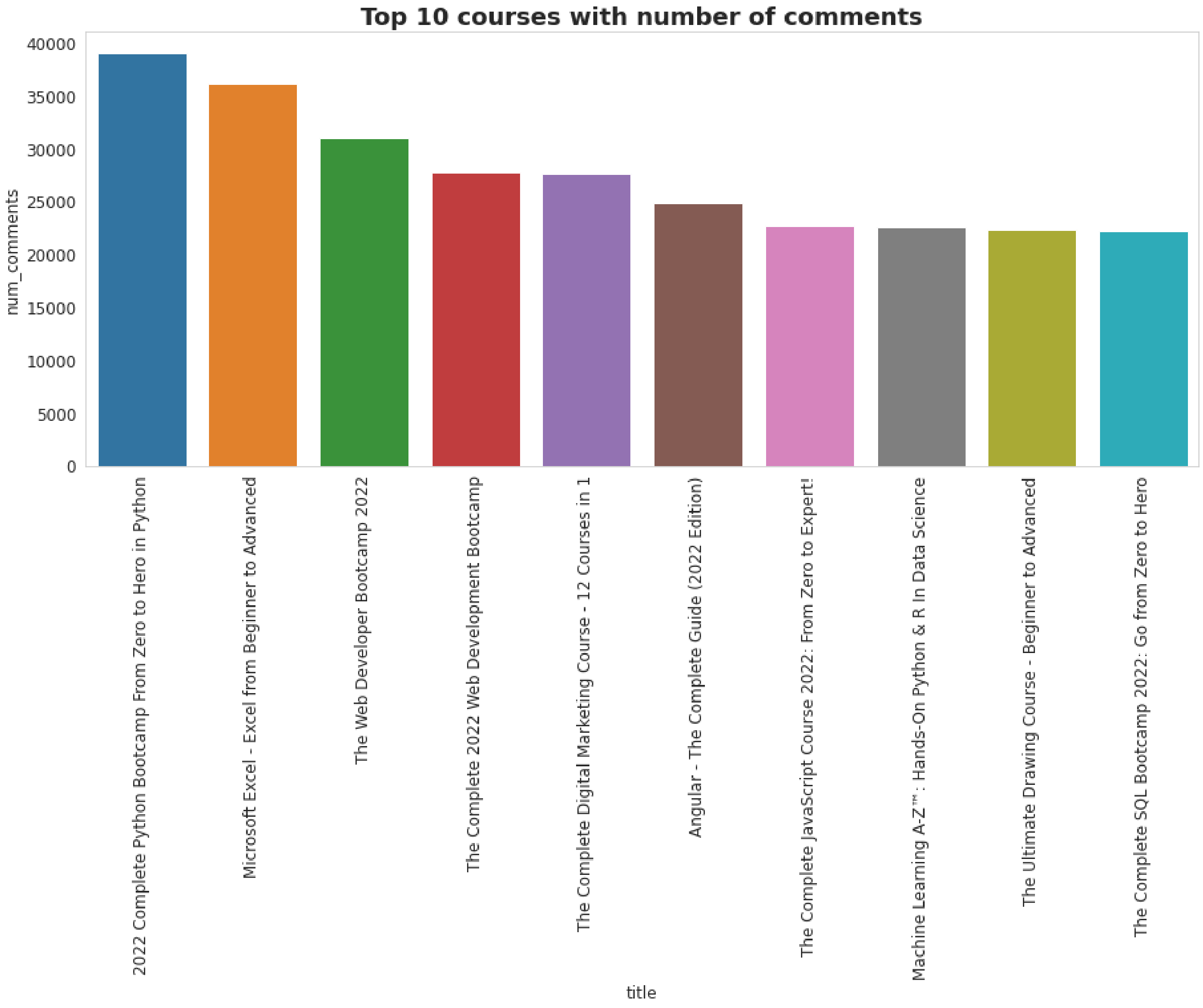
```
Top 10 instructors with their earnings in $:

instructor_name
Srinidhi Ranganathan     1,735,131,639.91
Learn Tech Plus          1,198,360,877.57
TJ Walker                1,171,615,935.48
Jose Portilla              818,904,784.12
YouAccel Training          801,222,163.35
Creative Online School     638,075,197.91
Robert (Bob) Steele        629,703,391.45
Kirill Eremenko            543,566,459.29
Joseph Delgadillo          543,485,234.46
365 Careers                 535,805,189.7
Name: earning, dtype: object
```

There are 72731 instructors on the Udemy platform.The total income of the instructors is calculatedby multiplying the 'price' and 'num_subscribers' columns. The above data shows the top 10 instructors with their income.There is no information is available on earnings calculation, discounts and coupons offered.

4]Which courses are the most active courses or we can say engaging courses by analyzing the comments of the courses?

```
Top_10_by_comments=udemy_courses_data[['title','instructor_name','num_comments']].sort_
```

```
fig,ax=plt.subplots(figsize=(15,6))
sns.barplot(data=Top_10_by_comments,x='title',y='num_comments')
plt.title("Top 10 courses with number of comments",weight='bold').set_fontsize(18)
plt.xticks(rotation=90);
```



```
Top_10_by_comments
```

	title	instructor_name	num_comments
10724	2022 Complete Python Bootcamp From Zero to Her...	Jose Portilla	39040.0
16288	Microsoft Excel - Excel from Beginner to Advanced	Kyle Pew	36101.0
12120	The Web Developer Bootcamp 2022	Colt Steele	31001.0
39113	The Complete 2022 Web Development Bootcamp	Dr. Angela Yu	27723.0
19303	The Complete Digital Marketing Course - 12 Cou...	Rob Percival	27540.0
15378	Angular - The Complete Guide (2022 Edition)	Maximilian Schwarzmüller	24886.0
17800	The Complete JavaScript Course 2022: From Zero...	Jonas Schmedtmann	22645.0
20252	Machine Learning A-Z™: Hands-On Python & R In ...	Kirill Eremenko	22567.0
18341	The Ultimate Drawing Course - Beginner to Adva...	Jaysen Batchelor	22341.0
15545	The Complete SQL Bootcamp 2022: Go from Zero t...	Jose Portilla	22141.0

In the list of top 10 there is one course on python(which was instructed by Jose Portilla) which is most interactive. There are 2 courses on Web development and 2 courses related to Machine Learning.

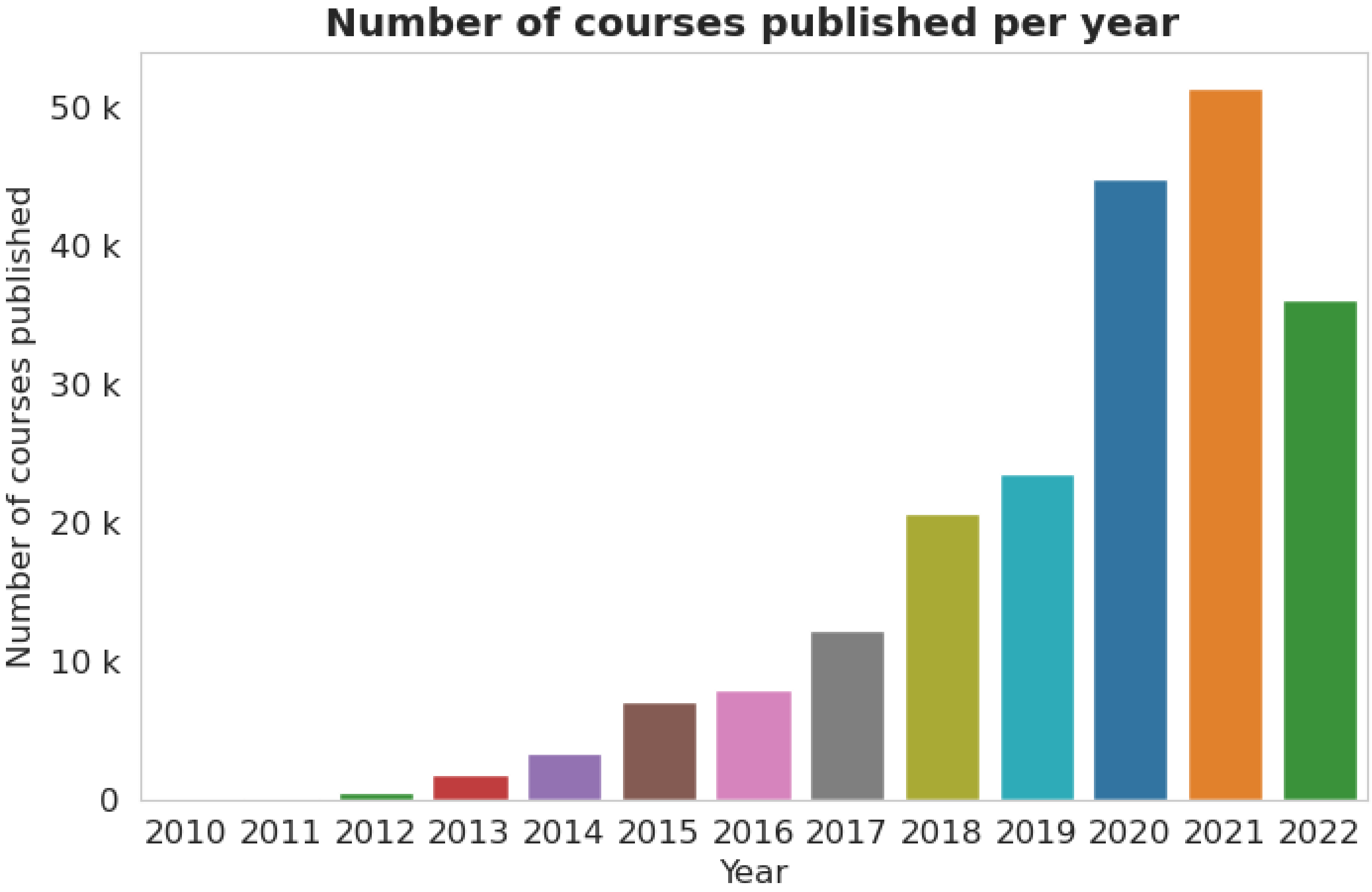
5]In which year the courses published are higher in number? Subscribers are also increased in that year? What about reviews and number of comments and lectures?

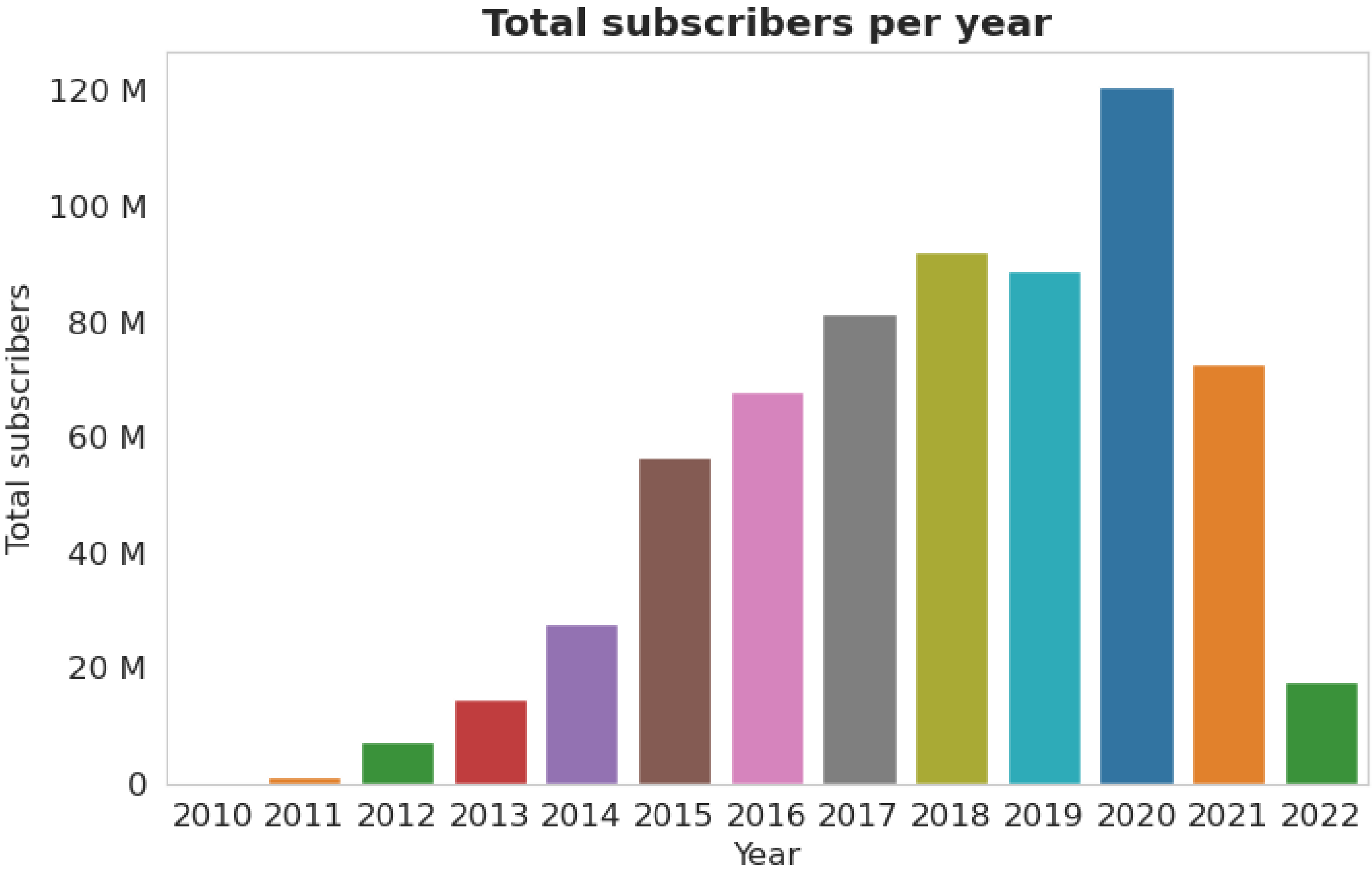
```
#To slove this question we have to do yearwise analysis of those column.
udemy_courses_data['published_year']=udemy_courses_data['published_time'].dt.year
#We need to group the column by year
year_wise_courses_count=udemy_courses_data['published_year'].value_counts()
columns_year_wise=udemy_courses_data.groupby('published_year')[['num_comments','num_sub
columns_year_wise=pd.pivot_table(udemy_courses_data,index='published_year',values=['id'
print('Year-wise count of number of courses and sum values of columns')
columns_year_wise
```

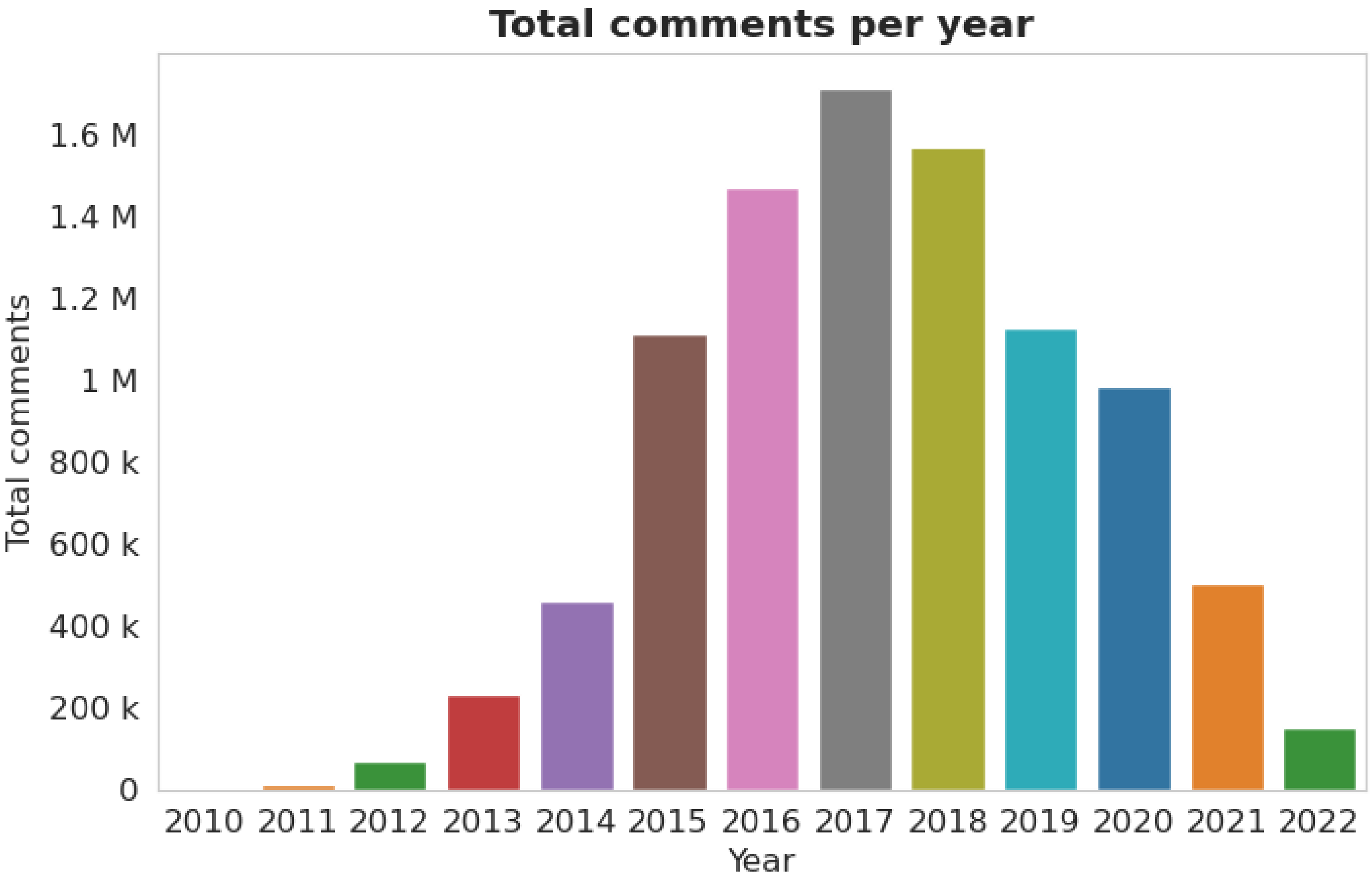
```
Year-wise count of number of courses and sum values of columns
```

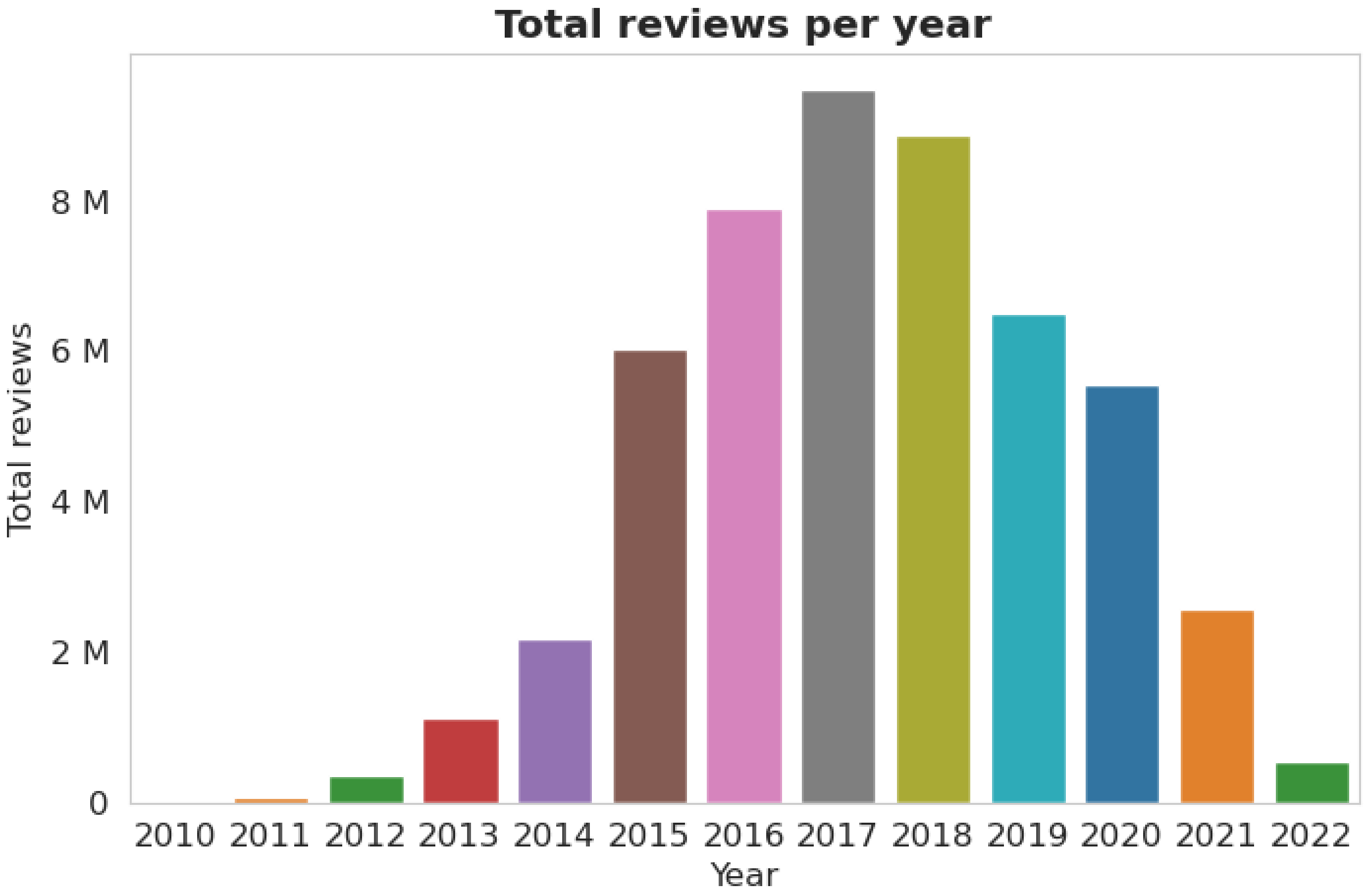
published_year	id	num_comments	num_lectures	num_reviews	num_subscribers
2010	4	408.0	68.0	1842.0	33727.0
2011	57	13170.0	4555.0	60041.0	1328159.0
2012	464	71397.0	25592.0	358467.0	7230026.0
2013	1778	233046.0	76847.0	1123207.0	14813615.0
2014	3404	461284.0	140262.0	2192152.0	27642396.0
2015	7102	1113930.0	298958.0	6042455.0	56527397.0
2016	7975	1470159.0	367520.0	7909710.0	68028864.0
2017	12258	1712435.0	561191.0	9486975.0	81376285.0
2018	20653	1568422.0	941243.0	8878711.0	92373156.0
2019	23529	1125888.0	974791.0	6498012.0	88915383.0
2020	44929	986528.0	1573159.0	5567350.0	120730813.0
2021	51457	504342.0	1618306.0	2587190.0	72812641.0
2022	36122	150718.0	1082949.0	544236.0	17726034.0

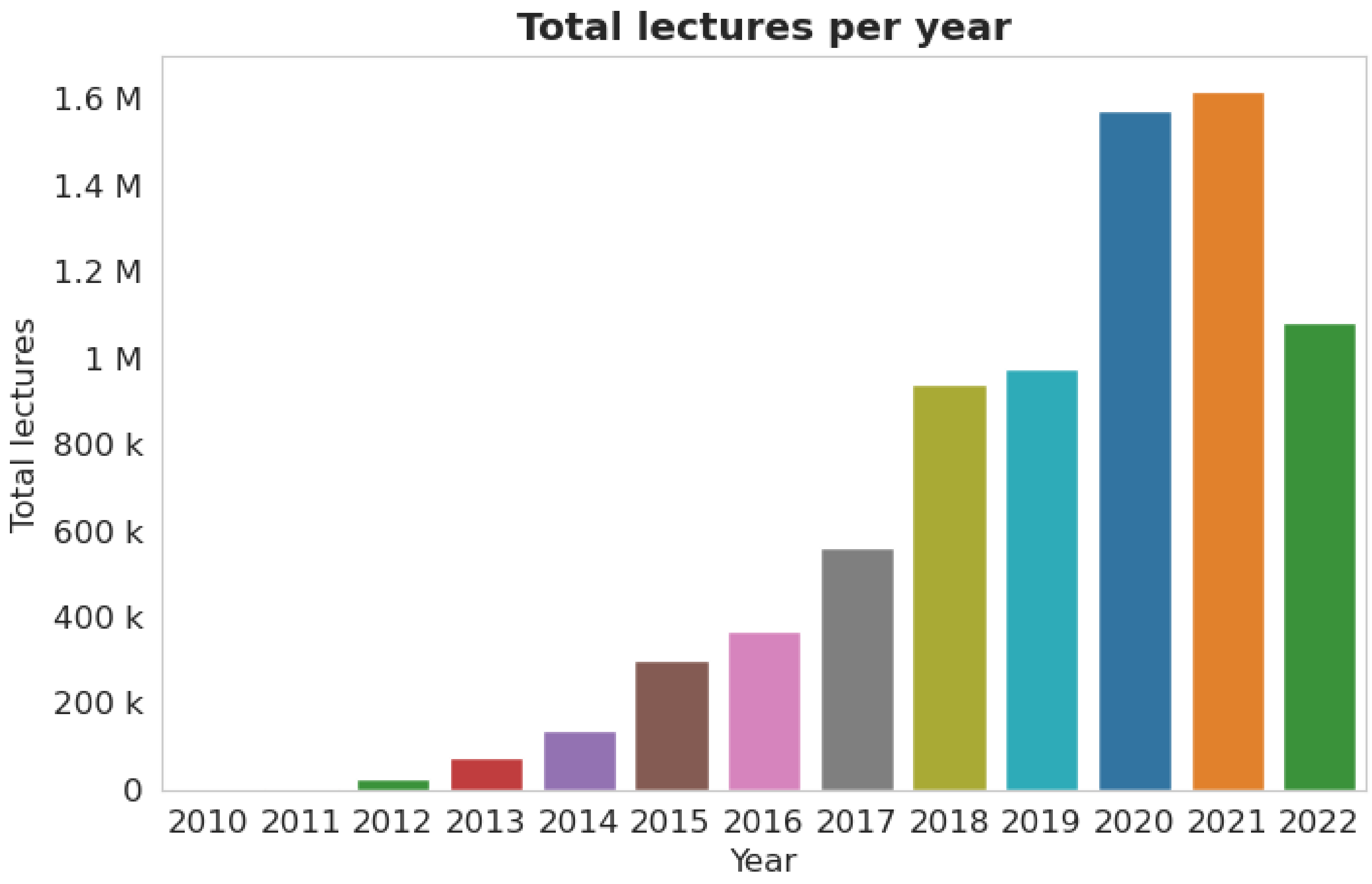
```
#let's plot this year wise number of courses published data
dict_columns=dict({'id':'Number of courses published','num_subscribers': 'Total subscri
for key, val in dict_columns.items():
    df=columns_year_wise
    fig,ax=plt.subplots(figsize=(8,5),dpi=100)
    sns.barplot(data=df,x=df.index,y=key,palette='tab10')
    ax.set_xlabel('Year')
    ax.set_ylabel(val)
    ax.set_title(f'{val} per year',weight='bold')
    ax.yaxis.set_major_formatter(ticker.EngFormatter());
```


Number of courses published per year
Number of courses published
0
10 k
20 k
30 k
40 k
50 k
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year


Total subscribers per year
Total subscribers
0
20 M
40 M
60 M
80 M
100 M
120 M
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year


Total comments per year
Total comments
1.6 M
1.4 M
1.2 M
1 M
800 k
600 k
400 k
200 k
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year


Total reviews per year
Total reviews
8 M
6 M
4 M
2 M
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year



Highest number of courses are published in year 2021, around 50 thousands courses are published.We can also say that after year 2019 number of courses started publishing on online platforms was increased. We all know the reason behind that is Covid-19 lock down.That's why the number of courses on udemy in 2019,2020,2021 are in higher number as compare previous years. In the last of 2021 and starting of year 2022 covid was recovered and improvement in health and lock downs also got over. So peoples again started offline learning. It might be happen in udemy that most of the instructor who were taking courses online again started offline teaching so the number of courses in 2022 was decreased as compared to 2021.

The number of subscribers in 2022 in compare with 2021 and 2020 are very less. The number of subscribers,number of comments,number of reviews are higher in 2017 year, It was because udemy was known to peoples in 2017.Udemy was become popular in 2017.

Inferences and conclusion

- **Udemy.**As the largest online learning provider, Udemy offers 209734 unique courses in 79 different languages which includes 59% from English,8.8% from Portugues,8.3% from Spanish ,etc.
- Out of 209734 courses, only 0.74% of data is missing.
- Before Udemy there was one site which was launshed by Bali,Octay Caglar and Gagan Biyani in early May 2010.
- 75% of the courses updates on and before March 17, 2022.
- There are 10.36% of the total courses i.e 21738 courses in Udemy which are available free of cost.
- The courses are covered under 13 categories and then these categories are further divided into 130 subcategories.
- There are 3818 unique topics under which different courses are offered.

- The category development containd the most number of courses(31643), followed by IT & Software(30479) and Teaching & Acdemics(26293).
- Out of top 10 courses in the list of highest number of subscribers, 2 courses are free.
- Development category has the highest number of subscribers(around 213M).and there over 15.7% of the courses on development which are free of cost.
- The category Music has the least number of subscribers(around 8.5M) with the instruments having the highest(around 3.9M) at the subcategory level.
- In the category of development, Web development field has most subscribers(around 76M).
- Most of the courses are priced between 0 and 200. Not all the courses are in the paid category. About 21738 courses are offered at no cost whereas 187996 courses have to be bought at a price.
- Over 500 courses having price greater than 900 Dollar.
- Out of total courses, 120k courses are offered in English language.
- In the list of topics, python has the number of subscribers(around 32M) in 2553 total courses.
- The course "2022 Complete Python Bootcamp From Zero to Hero" instructed by Jose Portilla has highest number of reviews as well as comments.We can also say that this course was most active course.
- The course 'Java Tutorial for Complete Beginners' instructed by John Purcell has highest number of subscribers.
- The course 'Chemistry for IIT JEE Main & Advanced, NEET, A...' instructed by Aman Saurav is most longest course.
- There are total 129 number of courses having price equal to 999.99 Dollar.
- There are 72731 instructors on the Udemy platform.From them 'Srinidhi Ranganathan' earned most around 1.7B Dollar.
- In udemy 6.86% of the courses (14392) has rating equal to 5.
- Dr Angela Yu course which is on "100 Days of Code: The Complete Python Pro Bootcamp for 2022" is the fastest growing course
- After 2019 courses were started publishing at high rate.It might be because of incresing the importance of online platform in Covid-19.
- But most subscribers,comment,reviews found in the year 2017.
- 2021 is the year when highest number of courses published(over 50k).

References and Future Work.

Resourses referred:

- Stack Overflow
- GeeksforGeeks
- Dataset-Kaggle
- Wikipedia
- Pandas

Projects referred:

- **US-accident-analysis**
- **General Elections Data Analysis**
- [** Udemy Course data and comments Analysis**(https://www.kaggle.com/code/hossaingh/udemy-courses-data-and-comments-analysis)

```
import jovian
jovian.commit()
```