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

Load Data

In [2]:
```
df=pd.read_csv("udemy_courses.csv")
```

Basic data exploration

In [3]:
```
df.head()
```

Out[3]:

	course_id	course_title	url	is_paid	price	num_subscribers	num_re
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	

In [4]:
```
df.level.value_counts()
```

Out[4]:
```
All Levels            1929
Beginner Level        1270
Intermediate Level     421
Expert Level            58
Name: level, dtype: int64
```

In [5]:
```
##value counts of all level of course
df.level.value_counts()
```

Out[5]:
```
All Levels            1929
Beginner Level        1270
Intermediate Level     421
Expert Level            58
Name: level, dtype: int64
```

Basic Data Cleaning

In [6]:
```
##Use LabelEncoder for encoading of is_paid column
from sklearn.preprocessing import LabelEncoder
labelencoder = LabelEncoder()
df['is_paid'] = labelencoder.fit_transform(df['is_paid'])
df.head()
```

Out[6]:

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

In [7]:
```
##Custom LabelEncoder for encoading level column
cleanup_nums = {"level":     {"All Levels": 0, "Beginner Level": 1,"Intermed
                              "Expert Level":3}}
df = df.replace(cleanup_nums)
df.head()
```

Out[7]:

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

In [8]:
```
df.isnull().sum()
##no null value in the data set
```

Out[8]:
```
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 [9]:
```
##showing datatypes of all column
df.dtypes
```

Out[9]:
```
course_id                int64
course_title            object
url                     object
is_paid                  int64
price                    int64
num_subscribers          int64
num_reviews              int64
num_lectures             int64
level                    int64
content_duration       float64
published_timestamp     object
subject                 object
dtype: object
```

In [10]:
```
df.is_paid.value_counts()
```

Out[10]:
```
1    3368
0     310
Name: is_paid, dtype: int64
```

In [11]:
```
df.subject.value_counts()
```

Out[11]:
```
Web Development        1200
Business Finance       1195
Musical Instruments     680
Graphic Design          603
Name: subject, dtype: int64
```

In [12]:
```
df.level.value_counts()
```

Out[12]:
```
0    1929
1    1270
2     421
3      58
Name: level, dtype: int64
```

All Tasks And Answers

In [13]:
```
## creating a new column name per_course_total_revenue
df["per_course_total_revenue"]=df["price"]*df["num_subscribers"]
```

In [14]:
```
1 df.head()
```

Out[14]:

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

In [15]:
```
1 ##for subject
2 ##Which subject got highest revenue?
3 df.loc[df['per_course_total_revenue'].idxmax()]
4 ##we see that webdevelopment got highest revenue at about rupees 24316800/-
```

Out[15]:
```
course_id                                                          625204
course_title                                       The Web Developer Bootcamp
url                          https://www.udemy.com/the-web-developer-bootcamp/
 (https://www.udemy.com/the-web-developer-bootcamp/)
is_paid                                                                 1
price                                                                 200
num_subscribers                                                    121584
num_reviews                                                         27445
num_lectures                                                          342
level                                                                   0
content_duration                                                       43
published_timestamp                                  2015-11-02T21:13:27Z
subject                                                   Web Development
per_course_total_revenue                                         24316800
Name: 3230, dtype: object
```

In [16]:
```
##See all column's unique values
def unique_value(df):
    for column in df:
        print(f'{column}:{df[column].unique()}')

##unique_value(df)
```

In [17]:
```
## All subjects name and their total courses counts
df.subject.value_counts()
```

Out[17]:
```
Web Development        1200
Business Finance       1195
Musical Instruments     680
Graphic Design          603
Name: subject, dtype: int64
```

Rvenue differences among all subjects with the help of pie chart

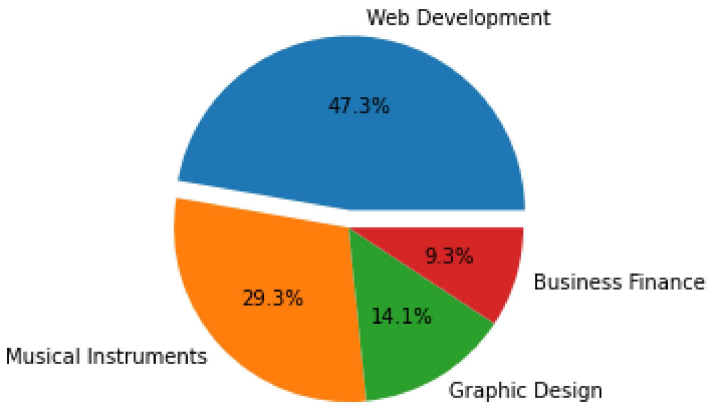
In [18]:
```
web=df[df['subject']=="Web Development"]
web_max=web.loc[web['per_course_total_revenue'].idxmax()]

Business_fin=df[df['subject']=="Business Finance"]
Business_fin_max=Business_fin.loc[Business_fin['per_course_total_revenue'].i

Musical_Instruments=df[df['subject']=="Musical Instruments"]
Musical_Instruments_max=Musical_Instruments.loc[Musical_Instruments['per_cou

Graphic_Design=df[df['subject']=="Graphic Design"]
Graphic_Design_max=Graphic_Design.loc[Graphic_Design['per_course_total_reven

##concatination of all this dataframes
revenue_dif=pd.concat([web_max,Business_fin_max,Musical_Instruments_max,Grap
revenue_dif=revenue_dif.T

```

In [19]:
```
##Showing the values from high to low
revenue_dif=revenue_dif.sort_values(by=['per_course_total_revenue'],ascendin
revenue_dif.head()
```

Out[19]:

	course_id	course_title	url	is_paid	price	num_subscribers	n
3230	625204	The Web Developer Bootcamp	https://www.udemy.com/the-web-developer-bootcamp/	1	200	121584	
1979	238934	Pianoforall - Incredible New Way To Learn Pian...	https://www.udemy.com/pianoforall-incredible-n...	1	200	75499	
1213	820194	Photoshop for Entrepreneurs - Design 11 Practi...	https://www.udemy.com/photoshop-for-entreprene...	1	200	36288	
40	648826	The Complete Financial Analyst Course 2017	https://www.udemy.com/the-complete-financial-a...	1	195	24481	

In [20]:
```
##  plotting pie chart  acoording to max
suject=['Web Development','Musical Instruments','Graphic Design','Business F
exp=[0.1,0,0,0]
plt.pie(revenue_dif.per_course_total_revenue,labels=suject,autopct='%2.1f%%'
plt.show()
```



In [21]:
```
w,b,m,g=df.subject.value_counts()
```

In [22]:
```
1 df.subject.value_counts()
```

Out[22]:
```
Web Development        1200
Business Finance       1195
Musical Instruments     680
Graphic Design          603
Name: subject, dtype: int64
```

In [23]:
```
1 df['per_course_total_revenue'].dtypes
```

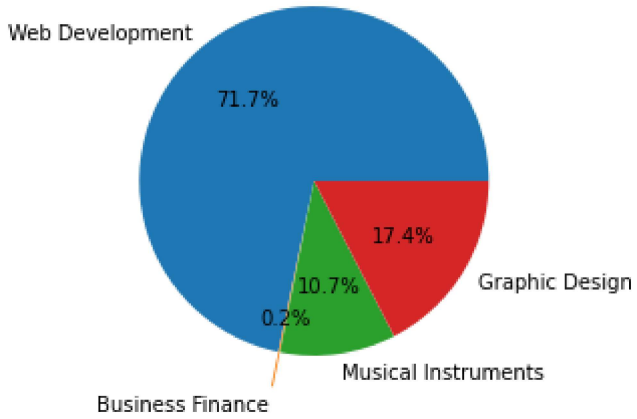
Out[23]:
```
dtype('int64')
```

In [24]:
```
1  ## For Average
2  web=df[df['subject']=='Web Development']['per_course_total_revenue'].sum()
3  web=web/w
4  bus=df[df['subject']=='Business Finance']['per_course_total_revenue'].sum()
5  bus=b
6  mus=df[df['subject']=='Musical Instruments']['per_course_total_revenue'].sum
7  mus=mus/m
8  grap=df[df['subject']=='Graphic Design']['per_course_total_revenue'].sum()
9  grap=grap/g
10 data=[web,bus,mus,grap]
11 subject=['Web Development','Business Finance','Musical Instruments','Graphic
12 index=[0,1,2,3]
13 q1=pd.DataFrame(data,index)
14 headers = ["Avg_revenue_per_subject"]
15 q1.columns = headers
16 q1['subject']=subject
17
18 q1.head()
```

Out[24]:

	Avg_revenue_per_subject	subject
0	525703.145833	Web Development
1	1195.000000	Business Finance
2	78469.198529	Musical Instruments
3	127666.948590	Graphic Design

In [25]:
```
##  plotting pie chart  acoording to Average
suject1=['Web Development','Business Finance','Musical Instruments','Graphic
exp=[0,0.2,0,0]
plt.pie(q1.Avg_revenue_per_subject  ,labels=suject1,autopct='%2.1f%%',explod
plt.show()
```



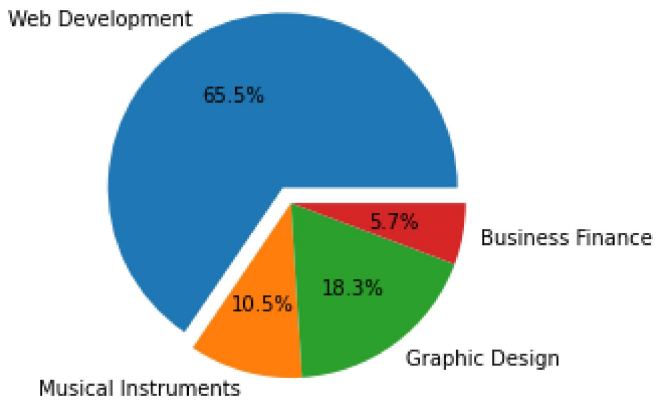
Which course is popular according to reviews and create a graph of that distribution

In [26]:
```
## Highest rating course
df.loc[df['num_reviews'].idxmax()]
```

Out[26]:
```
course_id                                                          625204
course_title                                         The Web Developer Bootcamp
url                          https://www.udemy.com/the-web-developer-bootcamp/
 (https://www.udemy.com/the-web-developer-bootcamp/)
is_paid                                                             1
price                                                             200
num_subscribers                                                121584
num_reviews                                                     27445
num_lectures                                                      342
level                                                               0
content_duration                                                   43
published_timestamp                                  2015-11-02T21:13:27Z
subject                                                  Web Development
per_course_total_revenue                                        24316800
Name: 3230, dtype: object
```

In [27]:
```
## for counting max
web=df[df['subject']=="Web Development"]
web_review_max=web.loc[web['num_reviews'].idxmax()]

Business_fin=df[df['subject']=="Business Finance"]
Business_review_fin_max=Business_fin.loc[Business_fin['num_reviews'].idxmax(

Musical_Instruments=df[df['subject']=="Musical Instruments"]
Musical_review_Instruments_max=Musical_Instruments.loc[Musical_Instruments['

Graphic_Design=df[df['subject']=="Graphic Design"]
Graphic_review_Design_max=Graphic_Design.loc[Graphic_Design['num_reviews'].i

##concatination of all this dataframes
review_df=pd.concat([web_review_max,Business_review_fin_max,Musical_review_I
review_df=review_df.T
```

In [28]:
```
##  plotting pie chart according to max
suject=['Web Development','Musical Instruments','Graphic Design','Business F
exp=[0.1,0,0,0]
plt.pie(review_df.num_reviews,labels=suject,autopct='%2.1f%%',explode=exp)
plt.show()
```

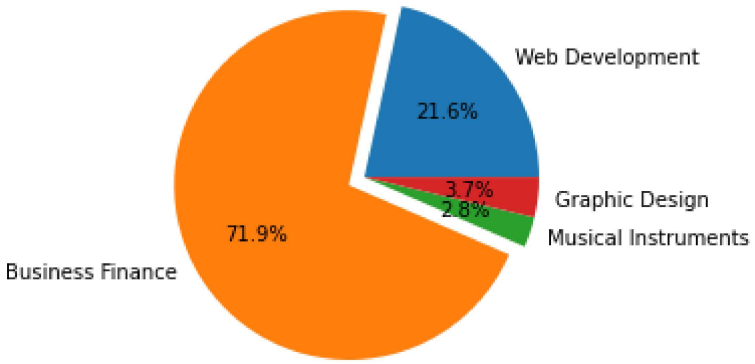


In [29]:
```
## For Average
web=df[df['subject']=='Web Development']['num_reviews'].sum()
web=web/w
bus=df[df['subject']=='Business Finance']['num_reviews'].sum()
bus=b
mus=df[df['subject']=='Musical Instruments']['num_reviews'].sum()
mus=mus/m
grap=df[df['subject']=='Graphic Design']['num_reviews'].sum()
grap=grap/g
data=[web,bus,mus,grap]
subject=['Web Development','Business Finance','Musical Instruments','Graphic
index=[0,1,2,3]
q2=pd.DataFrame(data,index)
headers = ["Avg_review_per_subject"]
q2.columns = headers
q2['subject']=subject

q2.head()
```

Out[29]:

	Avg_review_per_subject	subject
0	358.354167	Web Development
1	1195.000000	Business Finance
2	46.652941	Musical Instruments
3	61.475954	Graphic Design

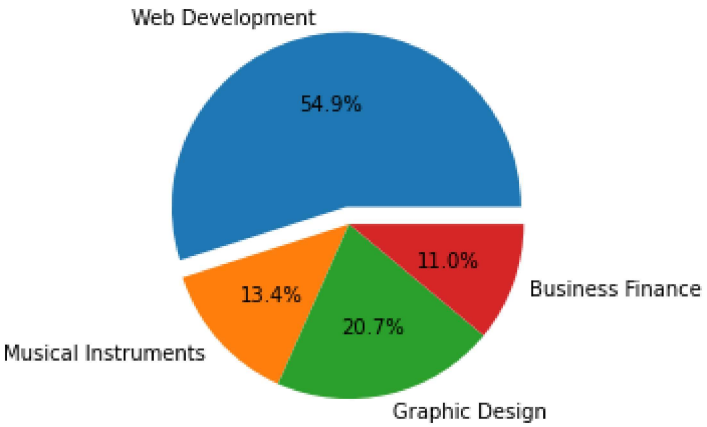
In [30]:
```
##  plotting pie chart  acoording to Average
suject1=['Web Development','Business Finance','Musical Instruments','Graphic
exp=[0,0.1,0,0]
plt.pie(q2.Avg_review_per_subject,labels=suject1,autopct='%2.1f%%',explode=e
plt.show()
```



Which course is popular according to subscribers and create a graph of that distribution

In [31]:
```
web=df[df['subject']=="Web Development"]
web_sub_max=web.loc[web['num_subscribers'].idxmax()]

Business_fin=df[df['subject']=="Business Finance"]
Business_sub_fin_max=Business_fin.loc[Business_fin['num_subscribers'].idxmax

Musical_Instruments=df[df['subject']=="Musical Instruments"]
Musical_sub_Instruments_max=Musical_Instruments.loc[Musical_Instruments['num

Graphic_Design=df[df['subject']=="Graphic Design"]
Graphic_sub_Design_max=Graphic_Design.loc[Graphic_Design['num_subscribers'].

##concatination of all this dataframes
review_sub=pd.concat([web_sub_max,Business_sub_fin_max,Musical_sub_Instrumen
review_sub=review_sub.T
```

In [32]:
```
##  plotting pie chart  acoording to max
suject=['Web Development','Musical Instruments','Graphic Design','Business F
exp=[0.1,0,0,0]
plt.pie(review_sub.num_subscribers,labels=suject,autopct='%2.1f%%',explode=e
plt.show()
```

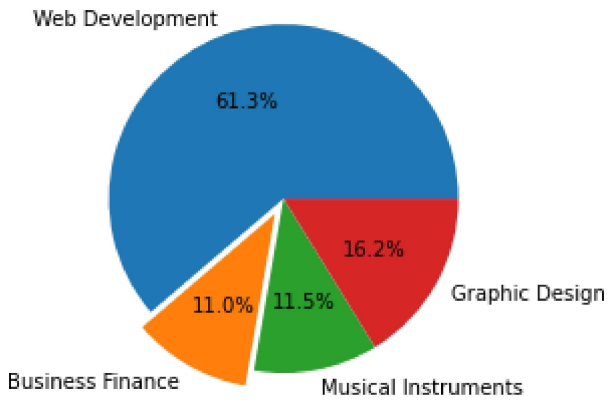


In [33]:
```
## For Average
web=df[df['subject']=='Web Development']['num_subscribers'].sum()
web=web/w
bus=df[df['subject']=='Business Finance']['num_subscribers'].sum()
bus=b
mus=df[df['subject']=='Musical Instruments']['num_subscribers'].sum()
mus=mus/m
grap=df[df['subject']=='Graphic Design']['num_subscribers'].sum()
grap=grap/g
data=[web,bus,mus,grap]
subject=['Web Development','Business Finance','Musical Instruments','Graphic
index=[0,1,2,3]
q3=pd.DataFrame(data,index)
headers = ["Avg_subscribers_per_subject"]
q3.columns = headers
q3['subject']=subject

q3.head()
```

Out[33]:

	Avg_subscribers_per_subject	subject
0	6650.476667	Web Development
1	1195.000000	Business Finance
2	1245.130882	Musical Instruments
3	1763.097844	Graphic Design

In [34]:
```
##  plotting pie chart  acoording to Average
suject1=['Web Development','Business Finance','Musical Instruments','Graphic
exp=[0,0.1,0,0]
plt.pie(q3.Avg_subscribers_per_subject,labels=suject1,autopct='%2.1f%%',expl
plt.show()
```



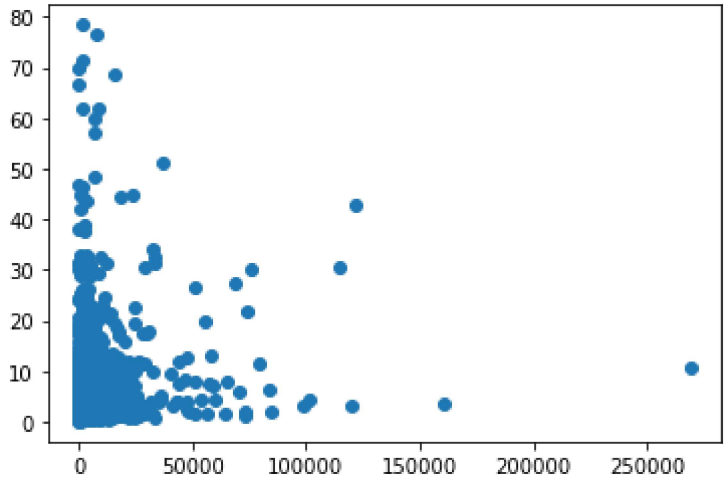
show the relation between course duration and subscribers

In [35]:
```
1 df.corr()
```

Out[35]:

	course_id	is_paid	price	num_subscribers	num_reviews	num_lec
course_id	1.000000	-0.013679	0.142319	-0.167856	-0.058550	-0.0
is_paid	-0.013679	1.000000	0.328513	-0.266159	-0.087471	0.1
price	0.142319	0.328513	1.000000	0.050769	0.113696	0.3
num_subscribers	-0.167856	-0.266159	0.050769	1.000000	0.649946	0.1
num_reviews	-0.058550	-0.087471	0.113696	0.649946	1.000000	0.2
num_lectures	-0.024646	0.112574	0.330160	0.157746	0.243029	1.0
level	0.078451	-0.033156	-0.073219	-0.062092	-0.055649	-0.1
content_duration	-0.057223	0.094417	0.293450	0.161839	0.228889	0.8
per_course_total_revenue	-0.053973	0.072902	0.346617	0.557175	0.769948	0.3

In [36]:
```
1 ##Correlation coefficient values less than +0.8 or greater than -0.8 are not
2 ##scatter plot to show the corelation
3 plt.scatter(df['num_subscribers'], df['content_duration'])
4 plt.show()
```



In [37]:
```
1 from scipy.stats import linregress
2 linregress(df['num_subscribers'], df['content_duration'])
3 ## here p alue is 5.2486796944911873e-23 which is less than 0.05 so relation
4 ## see the r value
```

Out[37]:
```
LinregressResult(slope=0.00010308643347981602, intercept=3.764934092352409, rva
lue=0.16183867741001334, pvalue=5.2486796944911873e-23, stderr=1.03673596696942
72e-05)
```

In [38]:
```
## Showing in different way
## p = population correlation coefficient (unknown)
## r = sample correlation coefficient (known; calculated from sample data)

np.corrcoef(df['num_subscribers'], df['content_duration'])

## Here r value is 0.16183868 which shows a positive corelation not more
```

Out[38]:
```
array([[1.        , 0.16183868],
       [0.16183868, 1.        ]])
```

In [39]:
```
##some notes

##If the p-value is less than the significance level (α = 0.05)
## Decision: Reject the null hypothesis.
##Conclusion: “There is sufficient evidence to conclude that there is a sign

## If the p-value is NOT less than the significance level (α = 0.05)
## Then Decision: DO NOT REJECT the null hypothesis.
##Conclusion: “There is insufficient evidence to conclude that there is a si


```

which course is popular according to level and create a graph of that distribution

In [40]:
```
pd.crosstab(df.level, df.subject)## it shows no of courses according to leve
```

Out[40]:

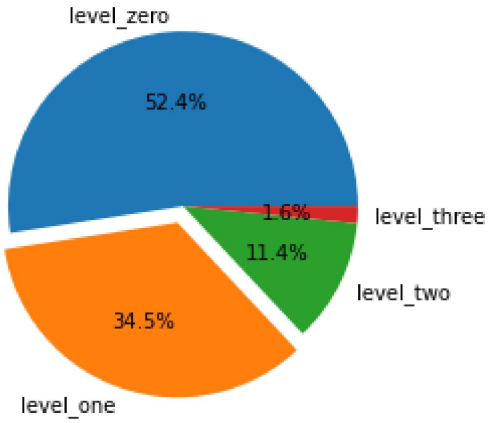
subject	Business Finance	Graphic Design	Musical Instruments	Web Development
level				
0	696	298	276	659
1	340	243	296	391
2	128	57	101	135
3	31	5	7	15

In [41]:
```
df.level.value_counts()
```

Out[41]:
```
0    1929
1    1270
2     421
3      58
Name: level, dtype: int64
```

In [42]:
```
df3=df.level.value_counts()
```

In [43]:
```
suject1=['level_zero','level_one','level_two','level_three']
exp=[0,0.1,0,0]
plt.pie(df3,labels=suject1,autopct='%2.1f%%',explode=exp)
plt.show()
```



Best unpaid course according to subscribers

In [44]:
```
df_unpaid=df[df['is_paid']==0]
df.loc[df_unpaid['num_subscribers'].idxmax()]
```

Out[44]:
```
course_id                                                        41295
course_title                                 Learn HTML5 Programming From Scratch
url                                https://www.udemy.com/learn-html5-programming-...
 (https://www.udemy.com/learn-html5-programming-...)
is_paid                                                              0
price                                                                0
num_subscribers                                                 268923
num_reviews                                                       8629
num_lectures                                                        45
level                                                                0
content_duration                                                  10.5
published_timestamp                                 2013-02-14T07:03:41Z
subject                                                  Web Development
per_course_total_revenue                                             0
Name: 2827, dtype: object
```

In [50]:
```
## Highest paid course according to subscribers
paid=df[df['is_paid']==1]
paid.loc[paid[paid['subject']=='Web Development']['per_course_total_revenue'
```

Out[50]:
```
course_id                                                          625204
course_title                                       The Web Developer Bootcamp
url                          https://www.udemy.com/the-web-developer-bootcamp/
 (https://www.udemy.com/the-web-developer-bootcamp/)
is_paid                                                                 1
price                                                                 200
num_subscribers                                                    121584
num_reviews                                                         27445
num_lectures                                                          342
level                                                                   0
content_duration                                                       43
published_timestamp                                  2015-11-02T21:13:27Z
subject                                                     Web Development
per_course_total_revenue                                           24316800
Name: 3230, dtype: object
```

In [ ]:
```

```