### EXPLORATARY DATA ANALYSIS ON UDEMY COURSES[¶](#EXPLORATARY-DATA-ANALYSIS-ON-UDEMY-COUR)

In [159]:

pwd

Out[159]:

'C:\\Users\\Administrator'

In [160]:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

LOAD DATA

In [162]:

df=pd.read\_csv("C:/Users/Administrator/Desktop/Udemy\_Courses.csv")

In [163]:

df.head()

Out[163]:

|  | **course\_id** | **course\_title** | **is\_paid** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** | **level** | **content\_duration** | **published\_timestamp** | **subject** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 288942 | 1 Piano Hand Coordination: Play 10th Ballad in... | True | 35.0 | 3137.0 | 18 | 68 | All Levels | 1.5 hours | 2014-09-18T05:07:05Z | Musical Instruments |
| **1** | 1170074 | 10 Hand Coordination - Transfer Chord Ballad 9... | True | 75.0 | 1593.0 | 1 | 41 | Intermediate Level | 1 hour | 2017-04-12T19:06:34Z | Musical Instruments |
| **2** | 1193886 | 12 Hand Coordination: Let your Hands dance wit... | True | 75.0 | 482.0 | 1 | 47 | Intermediate Level | 1.5 hours | 2017-04-26T18:34:57Z | Musical Instruments |
| **3** | 1116700 | 4 Piano Hand Coordination: Fun Piano Runs in 2... | True | 75.0 | 850.0 | 3 | 43 | Intermediate Level | 1 hour | 2017-02-21T23:48:18Z | Musical Instruments |
| **4** | 1120410 | 5 Piano Hand Coordination: Piano Runs in 2 B... | True | 75.0 | 940.0 | 3 | 32 | Intermediate Level | 37 mins | 2017-02-21T23:44:49Z | Musical Instruments |

### DATA CLEANING[¶](#DATA-CLEANING)

REPLACING NULL VALUES

In [166]:

df.isnull().sum()

Out[166]:

course\_id 0

course\_title 0

is\_paid 0

price 31

num\_subscribers 24

num\_reviews 0

num\_lectures 0

level 0

content\_duration 0

published\_timestamp 0

subject 0

dtype: int64

In [167]:

mean\_value=df['price'].mean()

In [168]:

df['price'].fillna(mean\_value, inplace=True)

In [169]:

df.isnull().sum()

Out[169]:

course\_id 0

course\_title 0

is\_paid 0

price 0

num\_subscribers 24

num\_reviews 0

num\_lectures 0

level 0

content\_duration 0

published\_timestamp 0

subject 0

dtype: int64

In [170]:

mean\_value\_subs=df['num\_subscribers'].mean()

In [171]:

mean\_value\_subs

Out[171]:

3202.8810825587752

In [172]:

df['num\_subscribers'].fillna(mean\_value\_subs, inplace=True)

In [173]:

df.isnull().sum()

Out[173]:

course\_id 0

course\_title 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

## OUTLIER DETECTING[¶](#OUTLIER-DETECTING)

In [175]:

df.describe()

Out[175]:

|  | **course\_id** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** |
| --- | --- | --- | --- | --- | --- |
| **count** | 3.682000e+03 | 3682.000000 | 3682.000000 | 3682.000000 | 3682.000000 |
| **mean** | 6.766121e+05 | 66.087373 | 3202.881083 | 156.093156 | 40.065182 |
| **std** | 3.436355e+05 | 60.722319 | 9492.532432 | 934.957204 | 50.373299 |
| **min** | 8.324000e+03 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 4.078430e+05 | 20.000000 | 113.250000 | 4.000000 | 15.000000 |
| **50%** | 6.885580e+05 | 45.000000 | 935.000000 | 18.000000 | 25.000000 |
| **75%** | 9.617515e+05 | 95.000000 | 2609.250000 | 67.000000 | 45.000000 |
| **max** | 1.282064e+06 | 200.000000 | 268923.000000 | 27445.000000 | 779.000000 |

In [176]:

plt.boxplot(df['num\_subscribers'])

Out[176]:

{'whiskers': [<matplotlib.lines.Line2D at 0x21ebe1c4fb0>,

<matplotlib.lines.Line2D at 0x21ebe1c7080>],

'caps': [<matplotlib.lines.Line2D at 0x21ebe1c4b90>,

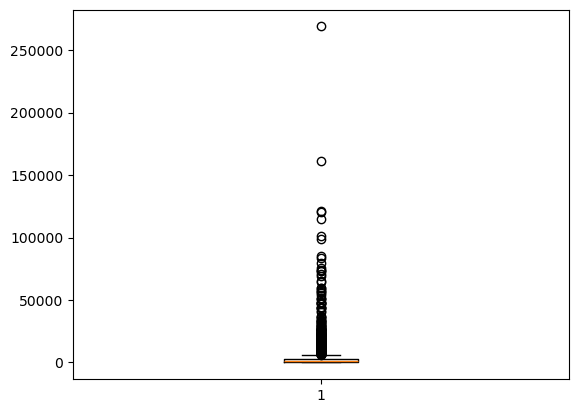
<matplotlib.lines.Line2D at 0x21ebe1c4530>],

'boxes': [<matplotlib.lines.Line2D at 0x21ebe1cc7d0>],

'medians': [<matplotlib.lines.Line2D at 0x21ebe1c6090>],

'fliers': [<matplotlib.lines.Line2D at 0x21ebe1c9910>],

'means': []}



In [212]:

Q1 = df['num\_subscribers'].quantile(0.25)

Q3 = df['num\_subscribers'].quantile(0.75)

IQR = Q3 - Q1

Lower\_bound = Q1 - 1.5 \* IQR

Upper\_bound = Q3 + 1.5 \* IQR

df['outliers'] =(df['num\_subscribers'] < Lower\_bound) | (df['num\_subscribers'] > Upper\_bound)

print(F"Q1:{Q1},Q3:{Q3},IQR:{IQR}")

print(F"Lower\_bound : {Lower\_bound},Upper\_bound : {Upper\_bound}")

Outliers = df[df['outliers']==True]

Q1:56.0,Q3:1375.5,IQR:1319.5

Lower\_bound : -1923.25,Upper\_bound : 3354.75

In [216]:

df\_cleaned = df.drop(Outliers.index)

In [179]:

df\_cleaned.head()

Out[179]:

|  | **course\_id** | **course\_title** | **is\_paid** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** | **level** | **content\_duration** | **published\_timestamp** | **subject** | **outliers** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 288942 | 1 Piano Hand Coordination: Play 10th Ballad in... | True | 35.0 | 3137.0 | 18 | 68 | All Levels | 1.5 hours | 2014-09-18T05:07:05Z | Musical Instruments | False |
| **1** | 1170074 | 10 Hand Coordination - Transfer Chord Ballad 9... | True | 75.0 | 1593.0 | 1 | 41 | Intermediate Level | 1 hour | 2017-04-12T19:06:34Z | Musical Instruments | False |
| **2** | 1193886 | 12 Hand Coordination: Let your Hands dance wit... | True | 75.0 | 482.0 | 1 | 47 | Intermediate Level | 1.5 hours | 2017-04-26T18:34:57Z | Musical Instruments | False |
| **3** | 1116700 | 4 Piano Hand Coordination: Fun Piano Runs in 2... | True | 75.0 | 850.0 | 3 | 43 | Intermediate Level | 1 hour | 2017-02-21T23:48:18Z | Musical Instruments | False |
| **4** | 1120410 | 5 Piano Hand Coordination: Piano Runs in 2 B... | True | 75.0 | 940.0 | 3 | 32 | Intermediate Level | 37 mins | 2017-02-21T23:44:49Z | Musical Instruments | False |

In [218]:

df=df\_cleaned

In [220]:

plt.boxplot(df['num\_subscribers'])

Out[220]:

{'whiskers': [<matplotlib.lines.Line2D at 0x21ebe5c2ae0>,

<matplotlib.lines.Line2D at 0x21ebe5c2510>],

'caps': [<matplotlib.lines.Line2D at 0x21ebe5c3110>,

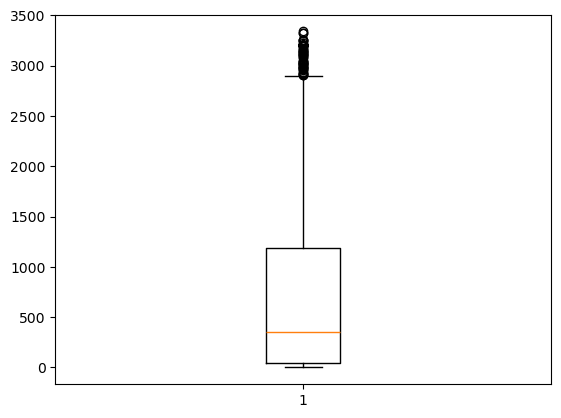
<matplotlib.lines.Line2D at 0x21ebe5c1e80>],

'boxes': [<matplotlib.lines.Line2D at 0x21ebe5c12e0>],

'medians': [<matplotlib.lines.Line2D at 0x21ebe5c2f90>],

'fliers': [<matplotlib.lines.Line2D at 0x21ebe5c0710>],

'means': []}



In [222]:

plt.boxplot(df['num\_reviews'])

Out[222]:

{'whiskers': [<matplotlib.lines.Line2D at 0x21ebe752660>,

<matplotlib.lines.Line2D at 0x21ebe752930>],

'caps': [<matplotlib.lines.Line2D at 0x21ebe752c00>,

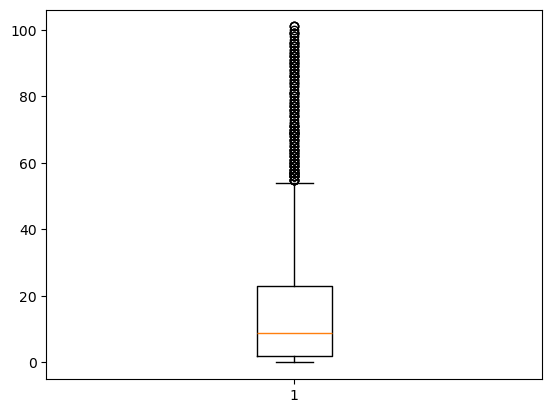
<matplotlib.lines.Line2D at 0x21ebe752ed0>],

'boxes': [<matplotlib.lines.Line2D at 0x21ebe1cd6d0>],

'medians': [<matplotlib.lines.Line2D at 0x21ebe7530b0>],

'fliers': [<matplotlib.lines.Line2D at 0x21ebe753380>],

'means': []}



In [224]:

Q1 = df['num\_reviews'].quantile(0.25)

Q3 = df['num\_reviews'].quantile(0.75)

IQR = Q3 - Q1

Lower\_bound = Q1 - 1.5 \* IQR

Upper\_bound = Q3 + 1.5 \* IQR

df['outliers'] =(df['num\_reviews'] < Lower\_bound) | (df['num\_reviews'] > Upper\_bound)

print(F"Q1:{Q1},Q3:{Q3},IQR:{IQR}")

print(F"Lower\_bound : {Lower\_bound},Upper\_bound : {Upper\_bound}")

Outliers = df[df['outliers']==True]

#print(F"\n Outliers:\n{Outliers}")

Q1:2.0,Q3:23.0,IQR:21.0

Lower\_bound : -29.5,Upper\_bound : 54.5

In [226]:

df\_cleaned = df.drop(Outliers.index)

In [228]:

df\_cleaned.head()

Out[228]:

|  | **course\_id** | **course\_title** | **is\_paid** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** | **level** | **content\_duration** | **published\_timestamp** | **subject** | **outliers** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 288942 | 1 Piano Hand Coordination: Play 10th Ballad in... | True | 35.0 | 3137.0 | 18 | 68 | All Levels | 1.5 hours | 2014-09-18T05:07:05Z | Musical Instruments | False |
| **1** | 1170074 | 10 Hand Coordination - Transfer Chord Ballad 9... | True | 75.0 | 1593.0 | 1 | 41 | Intermediate Level | 1 hour | 2017-04-12T19:06:34Z | Musical Instruments | False |
| **2** | 1193886 | 12 Hand Coordination: Let your Hands dance wit... | True | 75.0 | 482.0 | 1 | 47 | Intermediate Level | 1.5 hours | 2017-04-26T18:34:57Z | Musical Instruments | False |
| **3** | 1116700 | 4 Piano Hand Coordination: Fun Piano Runs in 2... | True | 75.0 | 850.0 | 3 | 43 | Intermediate Level | 1 hour | 2017-02-21T23:48:18Z | Musical Instruments | False |
| **4** | 1120410 | 5 Piano Hand Coordination: Piano Runs in 2 B... | True | 75.0 | 940.0 | 3 | 32 | Intermediate Level | 37 mins | 2017-02-21T23:44:49Z | Musical Instruments | False |

In [230]:

df=df\_cleaned

In [232]:

plt.boxplot(df['num\_reviews'])

Out[232]:

{'whiskers': [<matplotlib.lines.Line2D at 0x21ebea081d0>,

<matplotlib.lines.Line2D at 0x21ebea084d0>],

'caps': [<matplotlib.lines.Line2D at 0x21ebea087d0>,

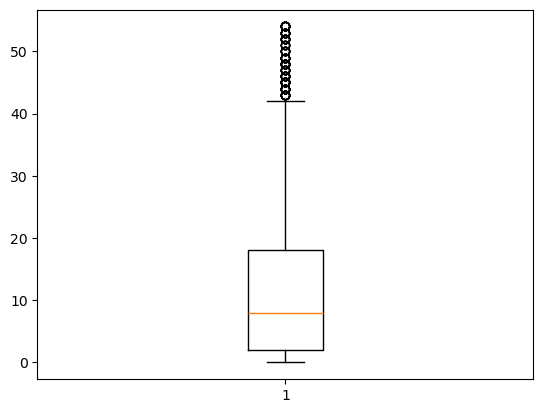
<matplotlib.lines.Line2D at 0x21ebea08aa0>],

'boxes': [<matplotlib.lines.Line2D at 0x21ebe787e60>],

'medians': [<matplotlib.lines.Line2D at 0x21ebea08c80>],

'fliers': [<matplotlib.lines.Line2D at 0x21ebea08f50>],

'means': []}



In [188]:

df.describe()

Out[188]:

|  | **course\_id** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** |
| --- | --- | --- | --- | --- | --- |
| **count** | 2.863000e+03 | 2863.000000 | 2863.000000 | 2863.000000 | 2863.000000 |
| **mean** | 7.015522e+05 | 61.319310 | 961.526854 | 19.405169 | 33.550122 |
| **std** | 3.429347e+05 | 55.905506 | 1232.353422 | 22.874128 | 38.369560 |
| **min** | 1.221400e+04 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 4.430030e+05 | 20.000000 | 59.500000 | 3.000000 | 14.000000 |
| **50%** | 7.201440e+05 | 40.000000 | 438.000000 | 10.000000 | 23.000000 |
| **75%** | 9.909740e+05 | 80.000000 | 1404.500000 | 28.000000 | 39.000000 |
| **max** | 1.282064e+06 | 200.000000 | 6315.000000 | 101.000000 | 462.000000 |

In [234]:

plt.boxplot(df['num\_lectures'])

Out[234]:

{'whiskers': [<matplotlib.lines.Line2D at 0x21ebea4a8d0>,

<matplotlib.lines.Line2D at 0x21ebea4ac00>],

'caps': [<matplotlib.lines.Line2D at 0x21ebea4ae40>,

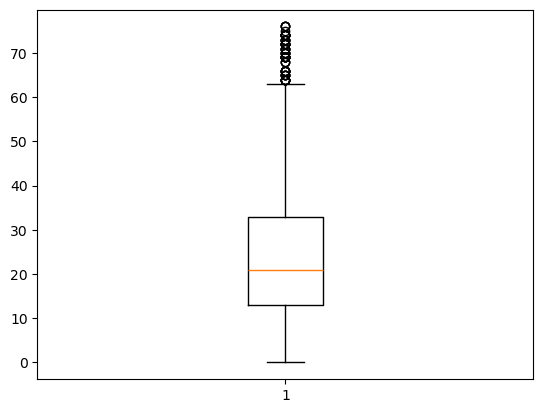
<matplotlib.lines.Line2D at 0x21ebea4b110>],

'boxes': [<matplotlib.lines.Line2D at 0x21ebea4a600>],

'medians': [<matplotlib.lines.Line2D at 0x21ebea4b050>],

'fliers': [<matplotlib.lines.Line2D at 0x21ebea4b290>],

'means': []}



In [236]:

Q1 = df['num\_lectures'].quantile(0.25)

Q3 = df['num\_lectures'].quantile(0.75)

IQR = Q3 - Q1

Lower\_bound = Q1 - 1.5 \* IQR

Upper\_bound = Q3 + 1.5 \* IQR

df['outliers'] =(df['num\_lectures'] < Lower\_bound) | (df['num\_lectures'] > Upper\_bound)

print(F"Q1:{Q1},Q3:{Q3},IQR:{IQR}")

print(F"Lower\_bound : {Lower\_bound},Upper\_bound : {Upper\_bound}")

Outliers = df[df['outliers']==True]

#print(F"\n Outliers:\n{Outliers}")

Q1:13.0,Q3:33.0,IQR:20.0

Lower\_bound : -17.0,Upper\_bound : 63.0

In [238]:

df\_cleaned = df.drop(Outliers.index)

In [240]:

df\_cleaned.head(5)

Out[240]:

|  | **course\_id** | **course\_title** | **is\_paid** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** | **level** | **content\_duration** | **published\_timestamp** | **subject** | **outliers** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 1170074 | 10 Hand Coordination - Transfer Chord Ballad 9... | True | 75.0 | 1593.0 | 1 | 41 | Intermediate Level | 1 hour | 2017-04-12T19:06:34Z | Musical Instruments | False |
| **2** | 1193886 | 12 Hand Coordination: Let your Hands dance wit... | True | 75.0 | 482.0 | 1 | 47 | Intermediate Level | 1.5 hours | 2017-04-26T18:34:57Z | Musical Instruments | False |
| **3** | 1116700 | 4 Piano Hand Coordination: Fun Piano Runs in 2... | True | 75.0 | 850.0 | 3 | 43 | Intermediate Level | 1 hour | 2017-02-21T23:48:18Z | Musical Instruments | False |
| **4** | 1120410 | 5 Piano Hand Coordination: Piano Runs in 2 B... | True | 75.0 | 940.0 | 3 | 32 | Intermediate Level | 37 mins | 2017-02-21T23:44:49Z | Musical Instruments | False |
| **5** | 1122832 | 6 Piano Hand Coordination: Play Open 10 Ballad... | True | 65.0 | 2015.0 | 3 | 21 | Intermediate Level | 44 mins | 2017-03-08T17:53:36Z | Musical Instruments | False |

In [242]:

df=df\_cleaned

In [244]:

plt.boxplot(df['num\_lectures'])

Out[244]:

{'whiskers': [<matplotlib.lines.Line2D at 0x21ebeac55b0>,

<matplotlib.lines.Line2D at 0x21ebeac5880>],

'caps': [<matplotlib.lines.Line2D at 0x21ebeac5b20>,

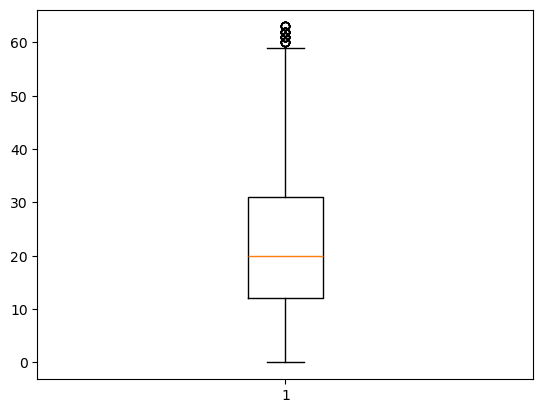
<matplotlib.lines.Line2D at 0x21ebeac5e20>],

'boxes': [<matplotlib.lines.Line2D at 0x21ebeac5370>],

'medians': [<matplotlib.lines.Line2D at 0x21ebeac5fd0>],

'fliers': [<matplotlib.lines.Line2D at 0x21ebeac6270>],

'means': []}



In [195]:

df.head()

Out[195]:

|  | **course\_id** | **course\_title** | **is\_paid** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** | **level** | **content\_duration** | **published\_timestamp** | **subject** | **outliers** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 288942 | 1 Piano Hand Coordination: Play 10th Ballad in... | True | 35.0 | 3137.0 | 18 | 68 | All Levels | 1.5 hours | 2014-09-18T05:07:05Z | Musical Instruments | False |
| **1** | 1170074 | 10 Hand Coordination - Transfer Chord Ballad 9... | True | 75.0 | 1593.0 | 1 | 41 | Intermediate Level | 1 hour | 2017-04-12T19:06:34Z | Musical Instruments | False |
| **2** | 1193886 | 12 Hand Coordination: Let your Hands dance wit... | True | 75.0 | 482.0 | 1 | 47 | Intermediate Level | 1.5 hours | 2017-04-26T18:34:57Z | Musical Instruments | False |
| **3** | 1116700 | 4 Piano Hand Coordination: Fun Piano Runs in 2... | True | 75.0 | 850.0 | 3 | 43 | Intermediate Level | 1 hour | 2017-02-21T23:48:18Z | Musical Instruments | False |
| **4** | 1120410 | 5 Piano Hand Coordination: Piano Runs in 2 B... | True | 75.0 | 940.0 | 3 | 32 | Intermediate Level | 37 mins | 2017-02-21T23:44:49Z | Musical Instruments | False |

In [196]:

df.describe()

Out[196]:

|  | **course\_id** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** |
| --- | --- | --- | --- | --- | --- |
| **count** | 2.640000e+03 | 2640.000000 | 2640.000000 | 2640.000000 | 2640.000000 |
| **mean** | 7.038863e+05 | 58.803858 | 959.807432 | 18.452273 | 25.365530 |
| **std** | 3.442906e+05 | 54.744054 | 1242.561449 | 22.116660 | 16.123743 |
| **min** | 1.221400e+04 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 4.451140e+05 | 20.000000 | 56.000000 | 3.000000 | 13.000000 |
| **50%** | 7.215100e+05 | 40.000000 | 426.000000 | 10.000000 | 21.000000 |
| **75%** | 9.939315e+05 | 70.000000 | 1375.500000 | 26.000000 | 34.000000 |
| **max** | 1.282064e+06 | 200.000000 | 6315.000000 | 101.000000 | 76.000000 |

In [197]:

df.head()

Out[197]:

|  | **course\_id** | **course\_title** | **is\_paid** | **price** | **num\_subscribers** | **num\_reviews** | **num\_lectures** | **level** | **content\_duration** | **published\_timestamp** | **subject** | **outliers** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 288942 | 1 Piano Hand Coordination: Play 10th Ballad in... | True | 35.0 | 3137.0 | 18 | 68 | All Levels | 1.5 hours | 2014-09-18T05:07:05Z | Musical Instruments | False |
| **1** | 1170074 | 10 Hand Coordination - Transfer Chord Ballad 9... | True | 75.0 | 1593.0 | 1 | 41 | Intermediate Level | 1 hour | 2017-04-12T19:06:34Z | Musical Instruments | False |
| **2** | 1193886 | 12 Hand Coordination: Let your Hands dance wit... | True | 75.0 | 482.0 | 1 | 47 | Intermediate Level | 1.5 hours | 2017-04-26T18:34:57Z | Musical Instruments | False |
| **3** | 1116700 | 4 Piano Hand Coordination: Fun Piano Runs in 2... | True | 75.0 | 850.0 | 3 | 43 | Intermediate Level | 1 hour | 2017-02-21T23:48:18Z | Musical Instruments | False |
| **4** | 1120410 | 5 Piano Hand Coordination: Piano Runs in 2 B... | True | 75.0 | 940.0 | 3 | 32 | Intermediate Level | 37 mins | 2017-02-21T23:44:49Z | Musical Instruments | False |

In [198]:

df.shape

Out[198]:

(2640, 12)

## Univariate Analysis[¶](#Univariate-Analysis)

In [200]:

plt.figure(figsize=(6,4))

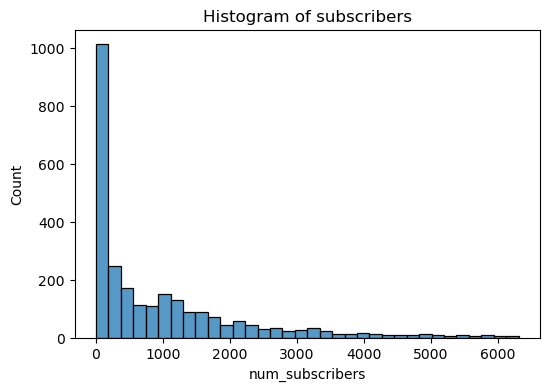
sns.histplot(df['num\_subscribers'])

bins=10,

kde=True

plt.title("Histogram of subscribers")

plt.show()



In [201]:

plt.figure(figsize=(4, 4))

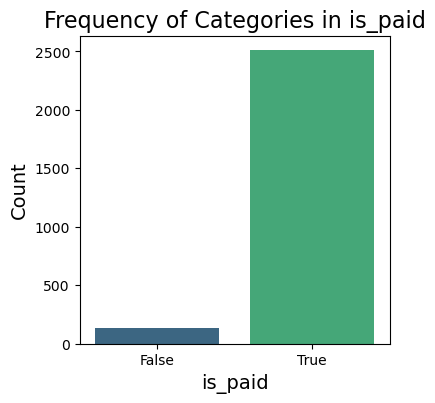
sns.countplot(x='is\_paid', data=df, palette='viridis')

plt.title(f'Frequency of Categories in {'is\_paid'}', fontsize=16)

plt.xlabel('is\_paid', fontsize=14)

plt.ylabel('Count', fontsize=14)

plt.show()



In [202]:

plt.figure(figsize=(7,4))

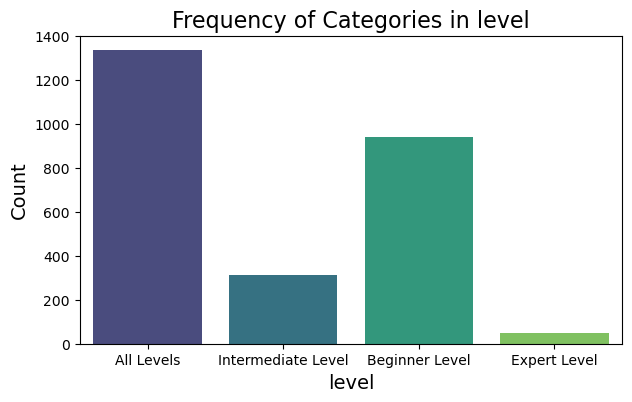
sns.countplot(x='level', data=df, palette='viridis')

plt.title(f'Frequency of Categories in {'level'}', fontsize=16)

plt.xlabel('level', fontsize=14)

plt.ylabel('Count', fontsize=14)

plt.show()



# Bivariate analysis[¶](#Bivariate-analysis)

In [204]:

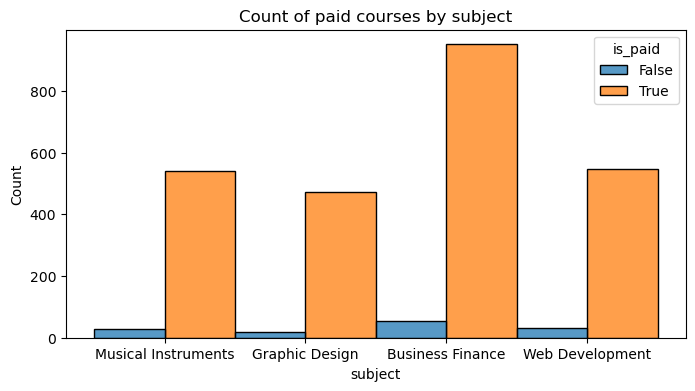
plt.figure(figsize=(8,4))

sns.histplot(x='subject', hue='is\_paid', data=df, stat="count", multiple="dodge")

plt.title('Count of paid courses by subject')

Out[204]:

Text(0.5, 1.0, 'Count of paid courses by subject')



In [255]:

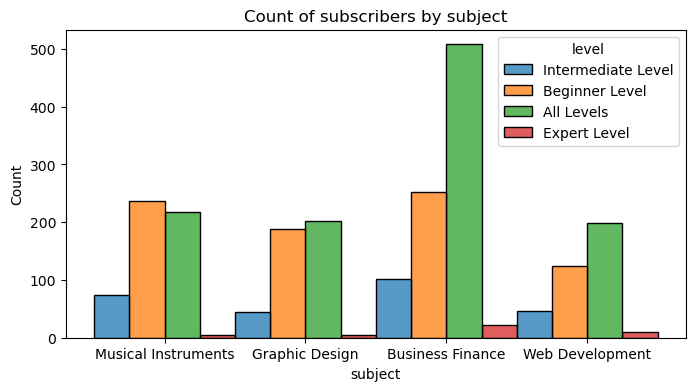
plt.figure(figsize=(8,4))

sns.histplot(x='subject', hue='level', data=df, stat="count", multiple="dodge")

plt.title('Count of subscribers by subject')

Out[255]:

Text(0.5, 1.0, 'Count of subscribers by subject')



## corelation matrix[¶](#corelation-matrix)

In [206]:

df\_numeric = df.select\_dtypes(include=['float64', 'int64'])

corr\_matrix = df\_numeric.corr()

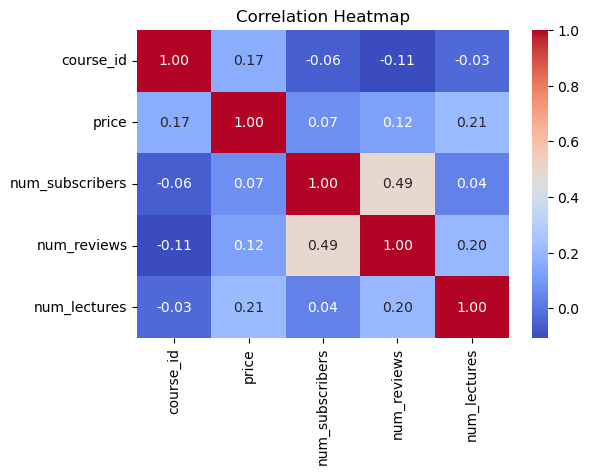
In [207]:

plt.figure(figsize=(6,4))

sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm', fmt=".2f")

plt.title('Correlation Heatmap')

plt.show()



In [ ]: