Python Pandas Project
Udemy online course dataset Analysis

By: Nirmal Gangadharan

INRODUCTION

This report presents the findings from an Exploratory Data Analysis (EDA) performed on the Udemy Online Courses dataset. The dataset provides insights into a wide range of courses available on Udemy, including attributes such as course titles, subjects, levels, prices, reviews, subscriber counts, and published timestamps.

About the dataset:

Columns : course_id, course_title, url, is_paid, price, num_subscribers, num_reviews, num_lectures, level, content_duration, published_timestamp, subject

Total number of columns = 12

Total number of rows = 3678

Data types:

course_id	int64
course_title	object
url	object
is_paid	bool
price	int64
num_subscribers	int64
num_reviews	int64
num_lectures	int64
level	object
content_duration	float64
published_timestamp	object
subject	object

Number of duplicate values = 6

Number of null values = 0

Original dataset:

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
0	1070968	Ultimate Investment Banking Course	https://www.udemy.com/ultimate-investment-bank...	True	200	2147	23	51	All Levels	1.5	2017-01-18T20:58:58Z	Business Finance
1	1113822	Complete GST Course & Certification - Grow You...	https://www.udemy.com/goods-and-services-tax/	True	75	2792	923	274	All Levels	39.0	2017-03-09T16:34:20Z	Business Finance
2	1006314	Financial Modeling for Business Analysts and C...	https://www.udemy.com/financial-modeling-for-b...	True	45	2174	74	51	Intermediate Level	2.5	2016-12-19T19:26:30Z	Business Finance
3	1210588	Beginner to Pro - Financial Analysis in Excel ...	https://www.udemy.com/complete-excel-finance-c...	True	95	2451	11	36	All Levels	3.0	2017-05-30T20:07:24Z	Business Finance
4	1011058	How To Maximize Your Profits Trading Options	https://www.udemy.com/how-to-maximize-your-pro...	True	200	1276	45	26	Intermediate Level	2.0	2016-12-13T14:57:18Z	Business Finance
...	...	...	...	...	...	...	...	...	...	...	...	...
3673	775618	Learn jQuery from Scratch - Master of JavaScri...	https://www.udemy.com/easy-jquery-for-beginner...	True	100	1040	14	21	All Levels	2.0	2016-06-14T17:36:46Z	Web Development
3674	1088178	How To Design A WordPress Website With No Codi...	https://www.udemy.com/how-to-make-a-wordpress-...	True	25	306	3	42	Beginner Level	3.5	2017-03-10T22:24:30Z	Web Development
3675	635248	Learn and Build using Polymer	https://www.udemy.com/learn-and-build-using-po...	True	40	513	169	48	All Levels	3.5	2015-12-30T16:41:42Z	Web Development
3676	905096	CSS Animations: Create Amazing Effects on Your...	https://www.udemy.com/css-animations-create-am...	True	50	300	31	38	All Levels	3.0	2016-08-11T19:06:15Z	Web Development
3677	297602	Using MODX CMS to Build Websites: A Beginner's...	https://www.udemy.com/using-modx-cms-to-build-...	True	45	901	36	20	Beginner Level	2.0	2014-09-28T19:51:11Z	Web Development

3678 rows × 12 columns

For the ease of analysis, the column ‘published_timestamp' was renamed to ‘uploaded_date’ and the values were converted to date.

Along with this, a new column ‘Year’ was feature engineered with reference to ‘uploaded_date’.

The updated dataset:

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	date_uploaded	subject	year
0	1070968	Ultimate Investment Banking Course	https://www.udemy.com/ultimate-investment-bank...	True	200	2147	23	51	All Levels	1.5	2017-01-18	Business Finance	2017
1	1113822	Complete GST Course & Certification - Grow You...	https://www.udemy.com/goods-and-services-tax/	True	75	2792	923	274	All Levels	39.0	2017-03-09	Business Finance	2017
2	1006314	Financial Modeling for Business Analysts and C...	https://www.udemy.com/financial-modeling-for-b...	True	45	2174	74	51	Intermediate Level	2.5	2016-12-19	Business Finance	2016
3	1210588	Beginner to Pro - Financial Analysis in Excel ...	https://www.udemy.com/complete-excel-finance-c...	True	95	2451	11	36	All Levels	3.0	2017-05-30	Business Finance	2017
4	1011058	How To Maximize Your Profits Trading Options	https://www.udemy.com/how-to-maximize-your-pro...	True	200	1276	45	26	Intermediate Level	2.0	2016-12-13	Business Finance	2016
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3666	775618	Learn jQuery from Scratch - Master of JavaScri...	https://www.udemy.com/easy-jquery-for-beginner...	True	100	1040	14	21	All Levels	2.0	2016-06-14	Web Development	2016
3667	1088178	How To Design A WordPress Website With No Codi...	https://www.udemy.com/how-to-make-a-wordpress-...	True	25	306	3	42	Beginner Level	3.5	2017-03-10	Web Development	2017
3668	635248	Learn and Build using Polymer	https://www.udemy.com/learn-and-build-using-po...	True	40	513	169	48	All Levels	3.5	2015-12-30	Web Development	2015
3669	905096	CSS Animations: Create Amazing Effects on Your...	https://www.udemy.com/css-animations-create-am...	True	50	300	31	38	All Levels	3.0	2016-08-11	Web Development	2016
3670	297602	Using MODX CMS to Build Websites: A Beginner's...	https://www.udemy.com/using-modx-cms-to-build-...	True	45	901	36	20	Beginner Level	2.0	2014-09-28	Web Development	2014

3671 rows × 13 columns

ANALYSIS

Understanding the data:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3678 entries, 0 to 3677
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   course_id            3678 non-null   int64
 1   course_title         3678 non-null   object
 2   url                  3678 non-null   object
 3   is_paid              3678 non-null   bool
 4   price                3678 non-null   int64
 5   num_subscribers      3678 non-null   int64
 6   num_reviews          3678 non-null   int64
 7   num_lectures         3678 non-null   int64
 8   level                3678 non-null   object
 9   content_duration     3678 non-null   float64
 10  published_timestamp  3678 non-null   object
 11  subject              3678 non-null   object
dtypes: bool(1), float64(1), int64(5), object(5)
```

```
df.describe()
```

	course_id	price	num_subscribers	num_reviews	num_lectures	content_duration
count	3.678000e+03	3678.000000	3678.000000	3678.000000	3678.000000	3678.000000
mean	6.759720e+05	66.049483	3197.150625	156.259108	40.108755	4.094517
std	3.432732e+05	61.005755	9504.117010	935.452044	50.383346	6.053840
min	8.324000e+03	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4.076925e+05	20.000000	111.000000	4.000000	15.000000	1.000000
50%	6.879170e+05	45.000000	911.500000	18.000000	25.000000	2.000000
75%	9.613555e+05	95.000000	2546.000000	67.000000	45.750000	4.500000
max	1.282064e+06	200.000000	268923.000000	27445.000000	779.000000	78.500000

```
df.shape
```

```
(3678, 12)
```

```
[8] df.head()
```

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
0	1070968	Ultimate Investment Banking Course	https://www.udemy.com/ultimate-investment-bank...	True	200	2147	23	51	All Levels	1.5	2017-01-18T20:58:58Z	Business Finance
1	1113822	Complete GST Course & Certification - Grow You...	https://www.udemy.com/goods-and-services-tax/	True	75	2792	923	274	All Levels	39.0	2017-03-09T16:34:20Z	Business Finance
2	1006314	Financial Modeling for Business Analysts and C...	https://www.udemy.com/financial-modeling-for-b...	True	45	2174	74	51	Intermediate Level	2.5	2016-12-19T19:26:30Z	Business Finance
3	1210588	Beginner to Pro - Financial Analysis in Excel ...	https://www.udemy.com/complete-excel-finance-c...	True	95	2451	11	36	All Levels	3.0	2017-05-30T20:07:24Z	Business Finance
4	1011058	How To Maximize Your Profits Trading Options	https://www.udemy.com/how-to-maximize-your-pro...	True	200	1276	45	26	Intermediate Level	2.0	2016-12-13T14:57:18Z	Business Finance

```
df.tail()
```

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
3673	775618	Learn jQuery from Scratch - Master of JavaScri...	https://www.udemy.com/easy-jquery-for-beginner...	True	100	1040	14	21	All Levels	2.0	2016-06-14T17:36:46Z	Web Development
3674	1088178	How To Design A WordPress Website With No Codi...	https://www.udemy.com/how-to-make-a-wordpress-...	True	25	306	3	42	Beginner Level	3.5	2017-03-10T22:24:30Z	Web Development
3675	635248	Learn and Build using Polymer	https://www.udemy.com/learn-and-build-using-po...	True	40	513	169	48	All Levels	3.5	2015-12-30T16:41:42Z	Web Development
3676	905096	CSS Animations: Create Amazing Effects on Your...	https://www.udemy.com/css-animations-create-am...	True	50	300	31	38	All Levels	3.0	2016-08-11T19:06:15Z	Web Development
3677	297602	Using MODX CMS to Build Websites: A Beginner's...	https://www.udemy.com/using-modx-cms-to-build-...	True	45	901	36	20	Beginner Level	2.0	2014-09-28T19:51:11Z	Web Development

Questions:

1) Courses with the highest content duration.

Query: df.sort_values(by='content_duration',ascending=False).head()

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level
1658	978576	The Complete Figure Drawing Course HD	https://www.udemy.com/the-complete-figure-draw...	True	50	1323	136	225	Beginner Level
3141	548278	The Complete Web Development Course - Build 1...	https://www.udemy.com/complete-web-development...	True	200	7501	1213	384	All Levels
561	375594	Financial Management - A Complete Study	https://www.udemy.com/financial-management-a-c...	True	190	1941	128	527	All Levels
874	167316	TRADER BOT: Introdução à Linguagem MQL5	https://www.udemy.com/intro-mql5/	True	20	209	33	33	All Levels
1214	62721	Anatomy for Figure Drawing: Mastering the Huma...	https://www.udemy.com/anatomy-for-figure-drawi...	True	95	15500	754	65	All Levels

2) Courses with lowest content duration.

Query: df.sort_values(by='content_duration').head()

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level
892	627332	Mutual Funds for Investors in Retirement Accounts	https://www.udemy.com/mutual-funds-for-investo...	True	20	0	0	0	All Levels
116	1191504	How to create a routine Trading	https://www.udemy.com/how-to-create-a-trading-...	True	25	307	8	5	All Levels
448	975074	ALGOTECH Hedge Fund Method for Stock Market Tr...	https://www.udemy.com/algotech-hedge-fund-meth...	True	20	605	19	4	All Levels
984	439210	Law Matters	https://www.udemy.com/law-matters/	True	20	327	1	12	Beginner Level
718	690546	UK Self Assessment Tax Return Filing Online	https://www.udemy.com/uk-tax-return/	True	20	23	7	6	All Levels

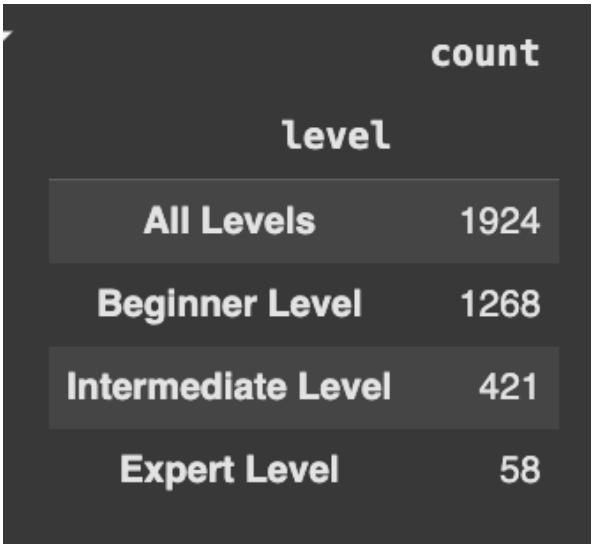
3) Highest priced courses

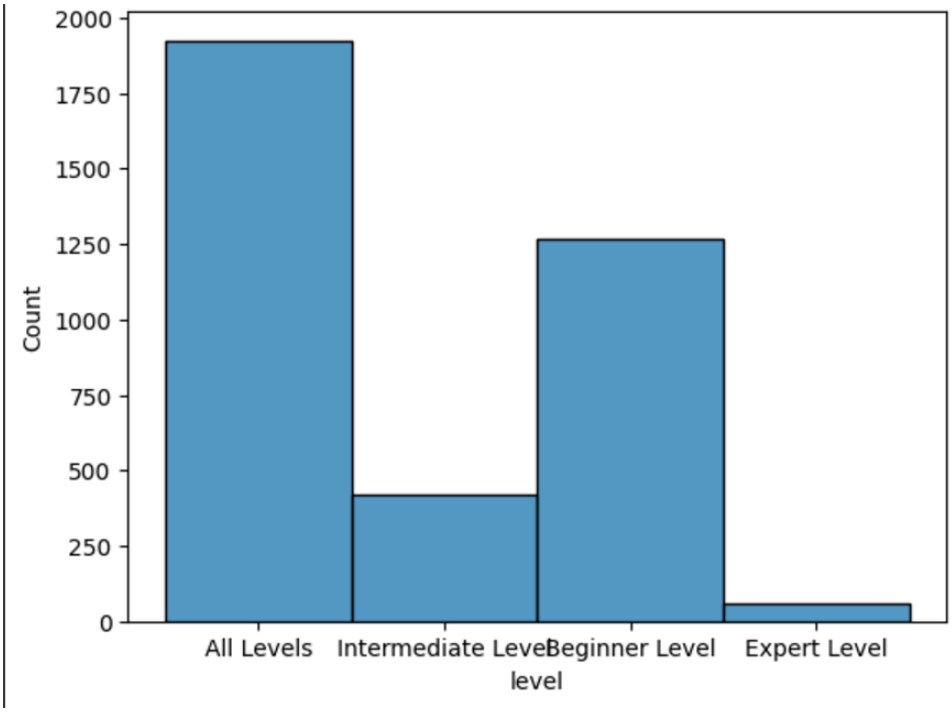
Query : df.sort_values(by='price',ascending=False).head()

	course_id	course_title	url	is_paid	price
0	1070968	Ultimate Investment Banking Course	https://www.udemy.com/ultimate-investment-bank...	True	200
2983	797040	Complete Guide to Front-End Web Development an...	https://www.udemy.com/complete-guide-to-front-...	True	200
3004	481696	Code & Grow Rich: Earn More As An Entrepreneu...	https://www.udemy.com/code-grow-rich-earn-more...	True	200
648	975414	Contango VXX - ETF Options Trading - Double Yo...	https://www.udemy.com/contango-vxx-trading-idi...	True	200
3016	1112604	Javascript Specialist	https://www.udemy.com/javascript-specialist/	True	200

4) Count of course of each levels

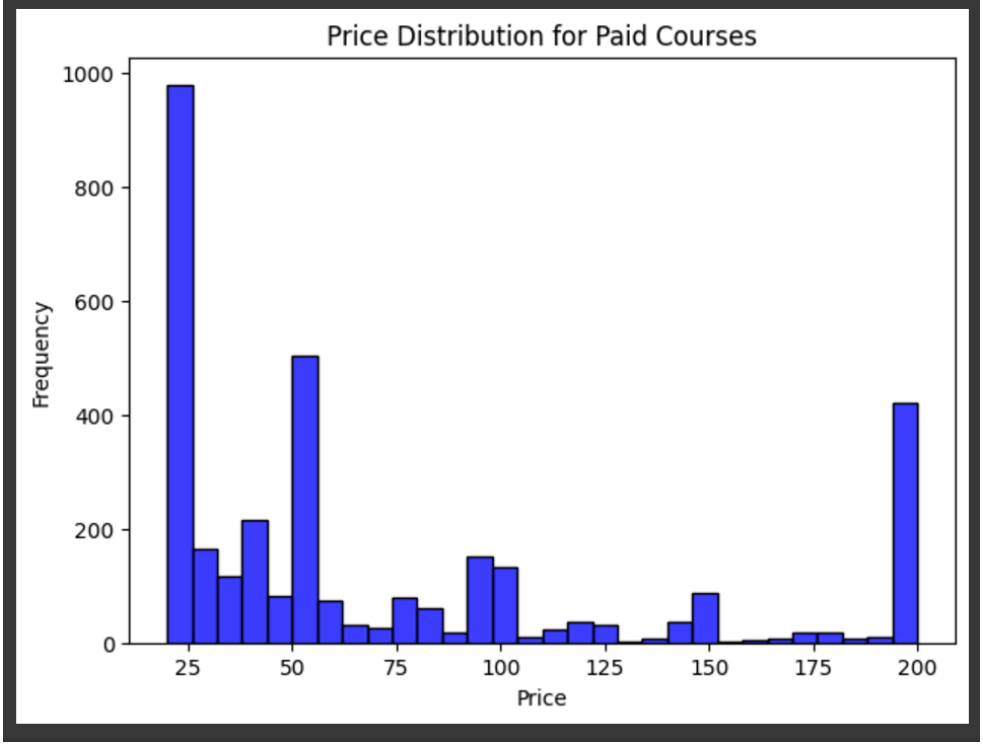
Query: df['level'].value_counts()





5) Price distribution for paid courses.

```
plt.figure(figsize=(7,5))
sns.histplot(df[df['is_paid']]['price'], bins=30,color="blue")
plt.title('Price Distribution for Paid Courses')
plt.xlabel("Price")
plt.ylabel("Frequency")
plt.show()
```



6) Courses with most number of subscribers

```
df.sort_values(by="num_subscribers",ascending=False).head()
```

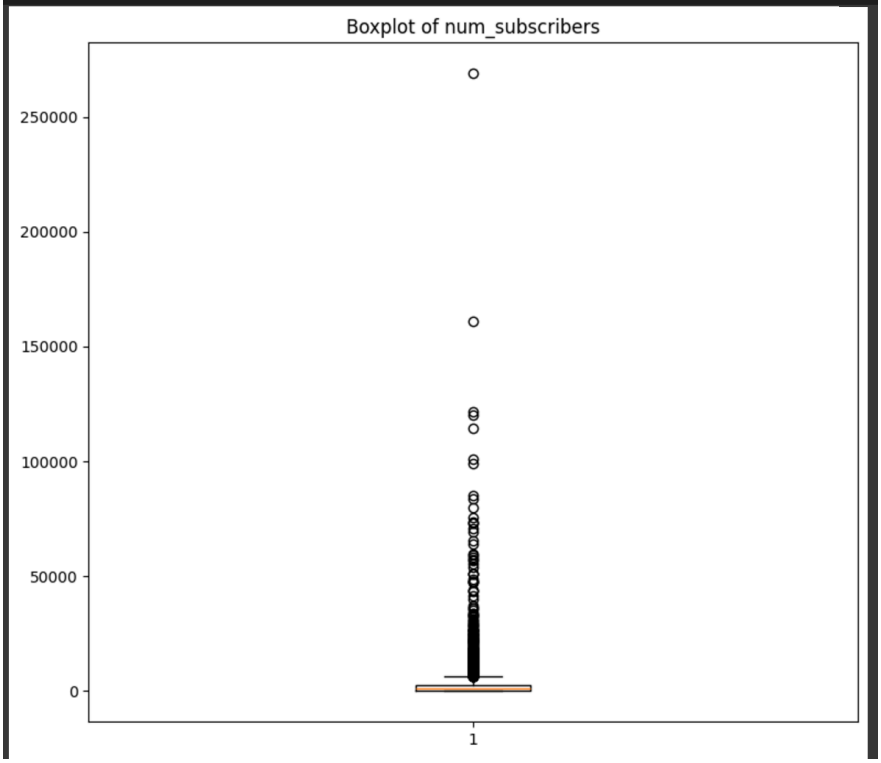
	index	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews
2820	2827	41295	Learn HTML5 Programming From Scratch	https://www.udemy.com/learn-html5-programming-...	False	0	268923	8629
3025	3032	59014	Coding for Entrepreneurs Basic	https://www.udemy.com/coding-for-entrepreneurs...	False	0	161029	279
3223	3230	625204	The Web Developer Bootcamp	https://www.udemy.com/the-web-developer-bootcamp/	True	200	121584	27445
2776	2783	173548	Build Your First Website in 1 Week with HTML5 ...	https://www.udemy.com/build-your-first-website...	False	0	120291	5924
3225	3232	764164	The Complete Web Developer Course 2.0	https://www.udemy.com/the-complete-web-develop...	True	200	114512	22412

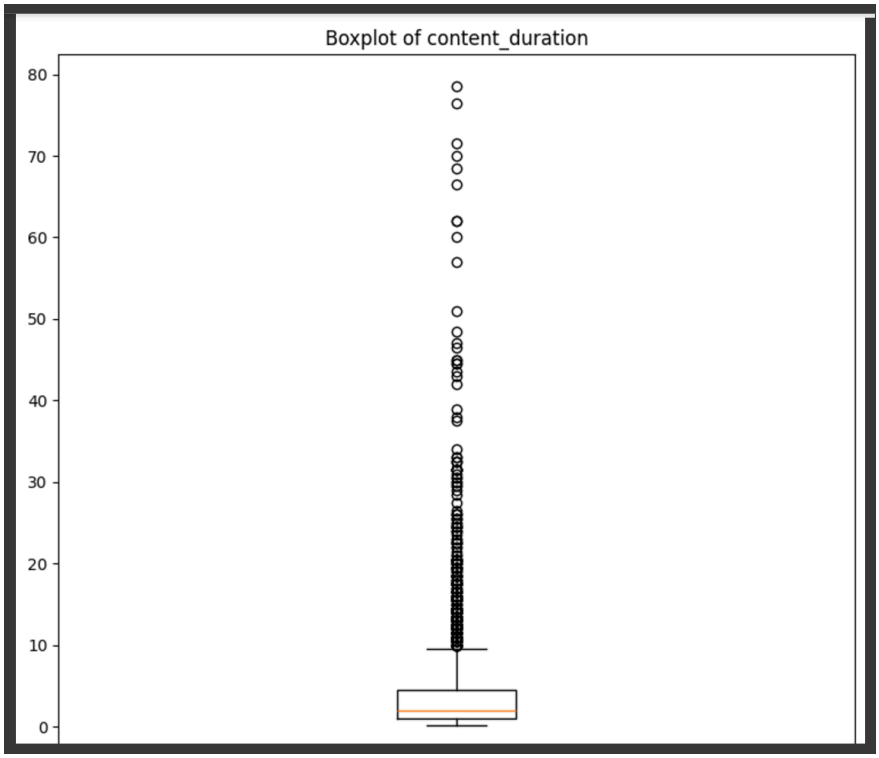
Analysis purely based on graphs:

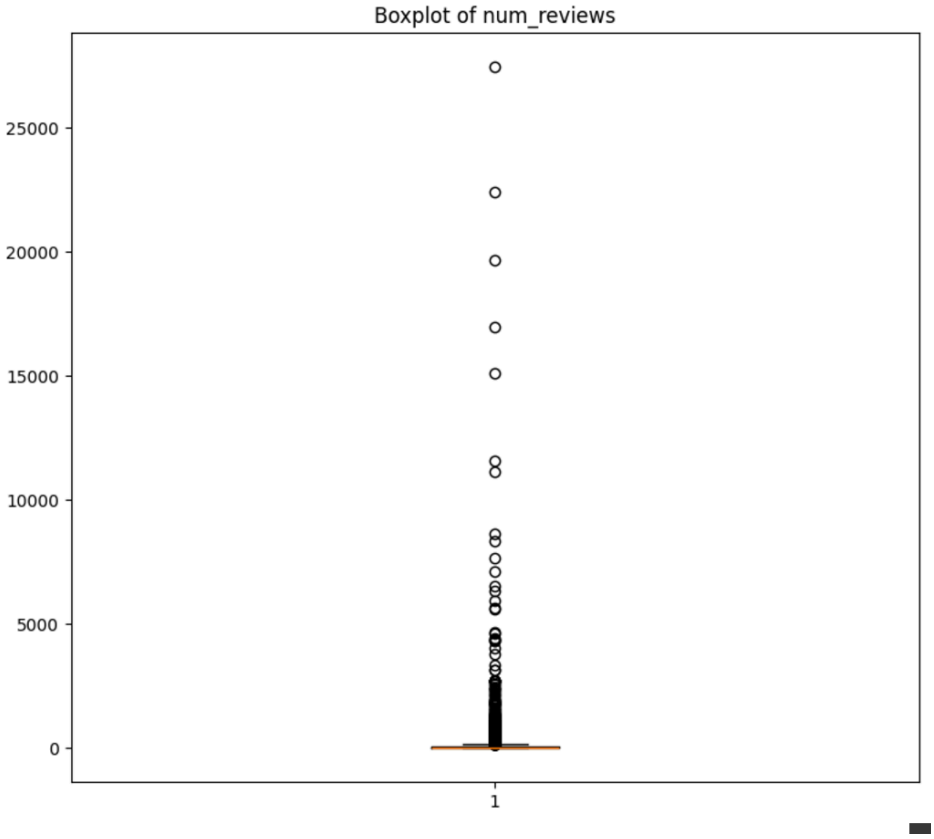
Box Plots were used to identify whether there is a outlier value among different column values

Some of them are:

```
for i in num1:
  plt.figure(figsize=(9,8))
  plt.boxplot(df[i])
  plt.title(f"Boxplot of {i}")
  plt.show()
```

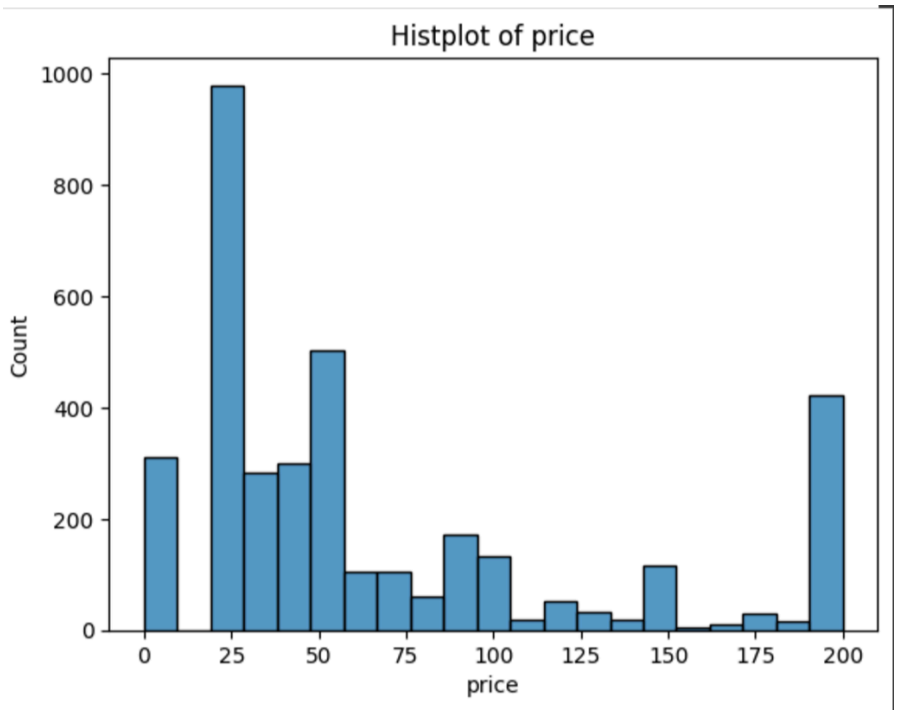




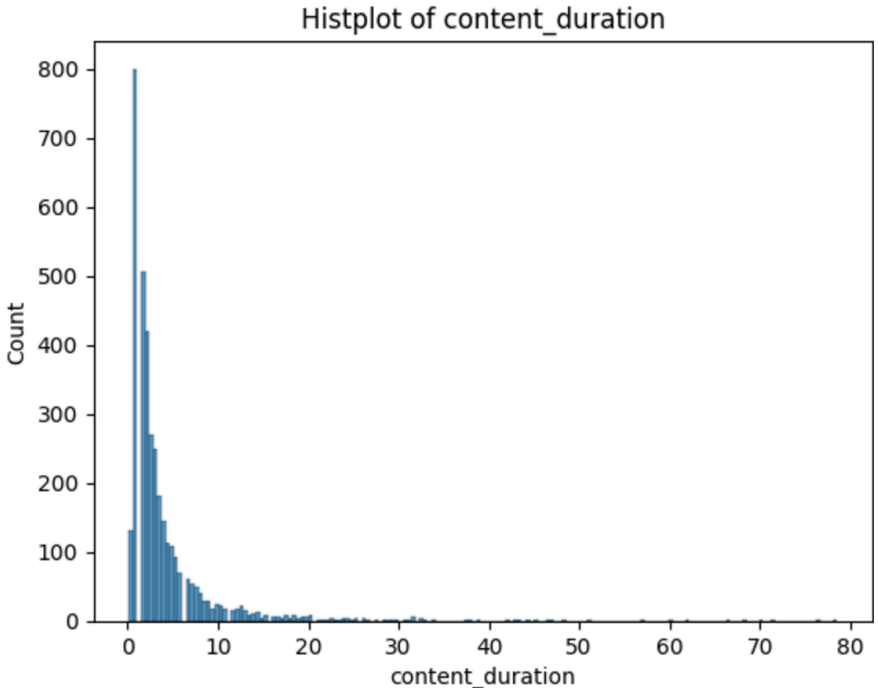


From these box plots no concerning outlier values were found and hence moved on with further analysis.

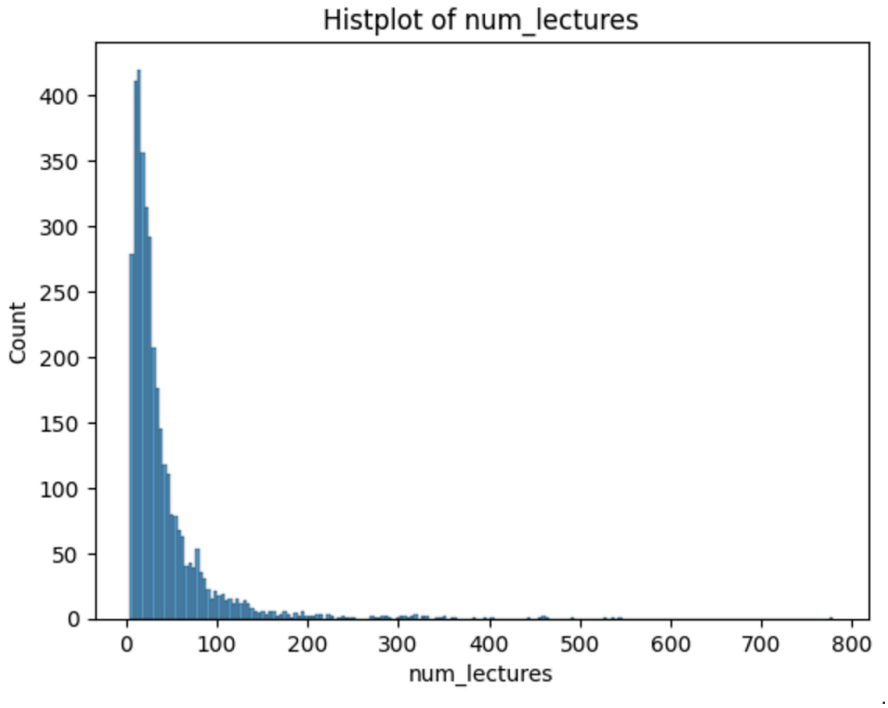
Use of Hist plot:



Here using this hist plot we can understand that most of the courses are offered at a price 25, followed by 50 and 200.



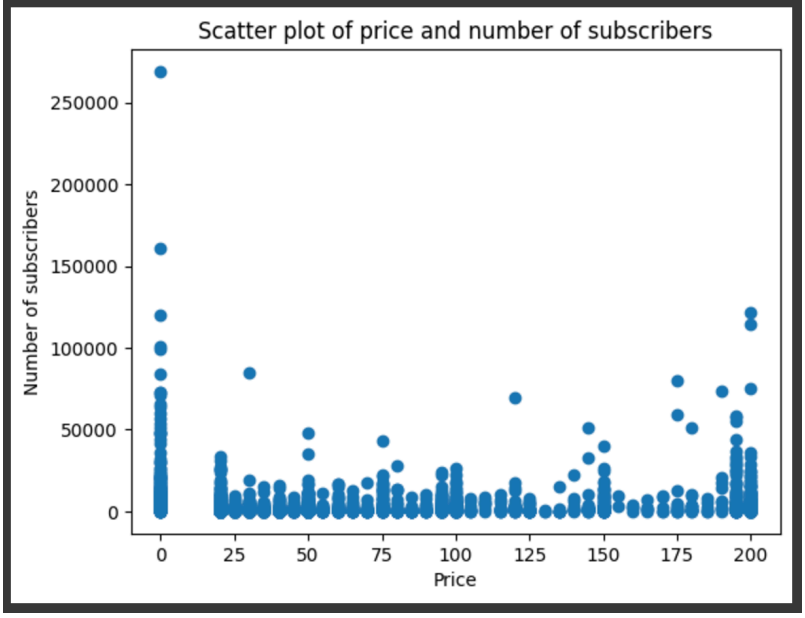
The above hist plot shows that, most of courses have content duration between 0 to almost 5 hours.



This hist plot shows that, most of the courses offered contains an average of 1 to 50 lectures in them, and the longest number of lectures span over 500.

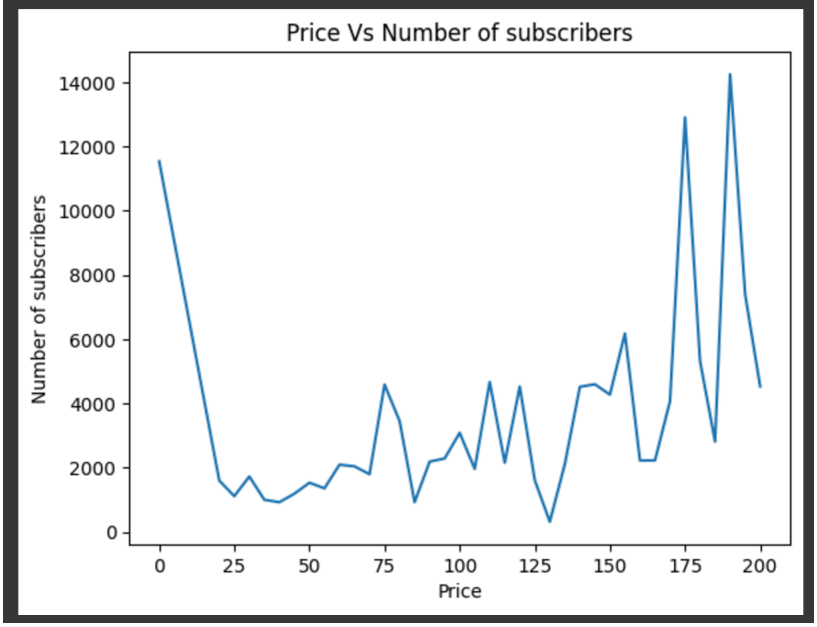
Scatter Plot:

```
plt.scatter(df['price'],df['num_subscribers'])
plt.title("Scatter plot of price and number of subscribers")
plt.xlabel("Price")
plt.ylabel("Number of subscribers")
plt.show()
```



This scatter plot helps to study the relationship between the price and the number of subscribers. Here, we can see that most people prefer free courses and also we can see that there is a strong demand for courses offered at price nearly 200.

```
var1=df[['price','num_subscribers']].groupby('price').mean()
plt.plot(var1.index,var1)
plt.title("Price Vs Number of subscribers")
plt.xlabel("Price")
plt.ylabel("Number of subscribers")
plt.show()
```

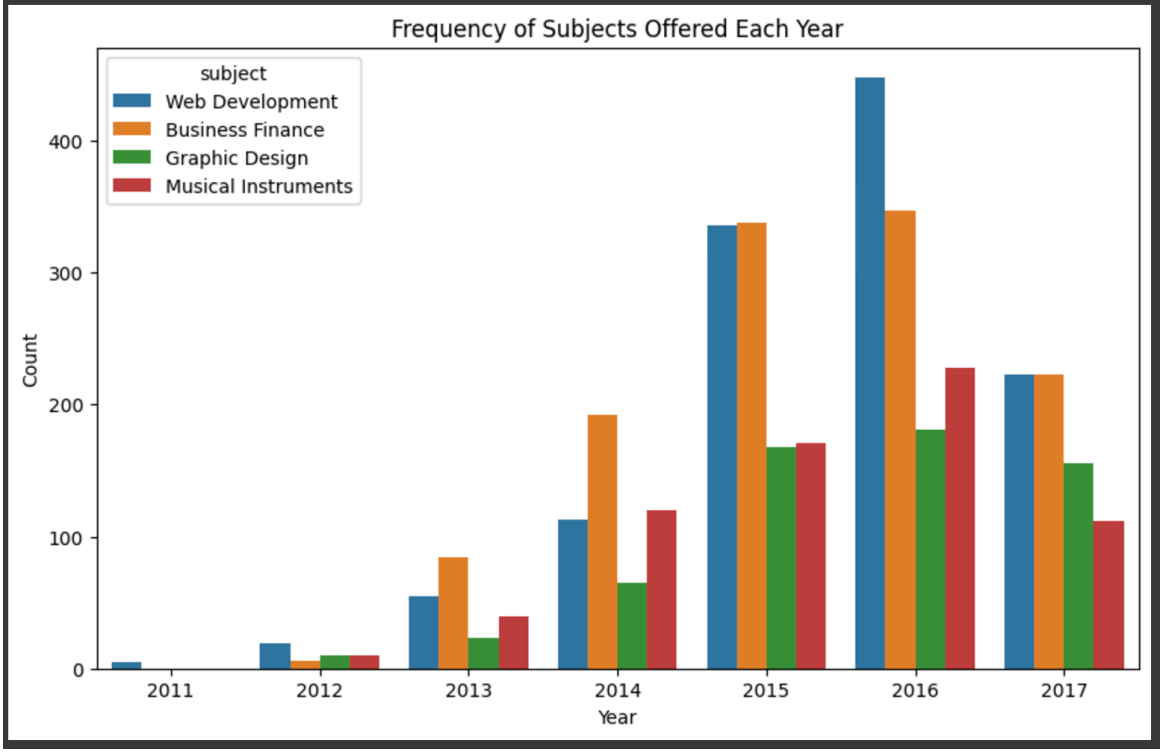


This line plot helps to understand the pattern much better.

Now let us look at the following Bar graph, In this graph we can see that courses offered are based on mainly four subjects(Business Finance, Graphic Design, Musical Instruments and Web development)

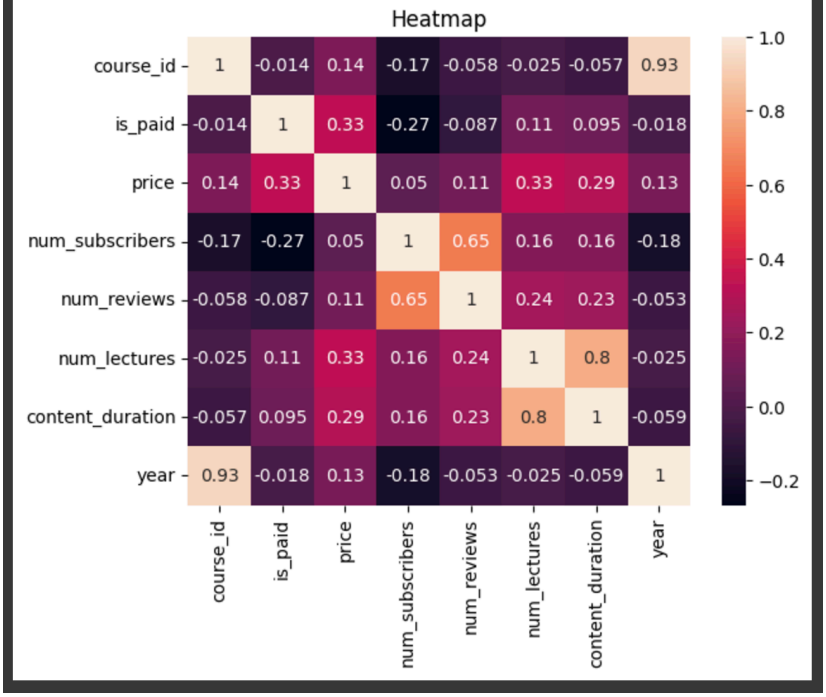
We can see the trend on each year, which course is mostly added on each year and we can also see that the year 2016 all the courses has been reached its maximum value.

```
var4 = df.groupby(['subject', 'year']).size().reset_index(name='count')
plt.figure(figsize=(10, 6))
sns.barplot(data=var4, x='year', y='count', hue='subject')
plt.title("Frequency of Subjects Offered Each Year")
plt.xlabel("Year")
plt.ylabel("Count")
plt.show()
```



Heat Map:

```
var3=df.corr(numeric_only=True)
plt.title('Heatmap')
sns.heatmap(var3,annot=True)
plt.show()
```



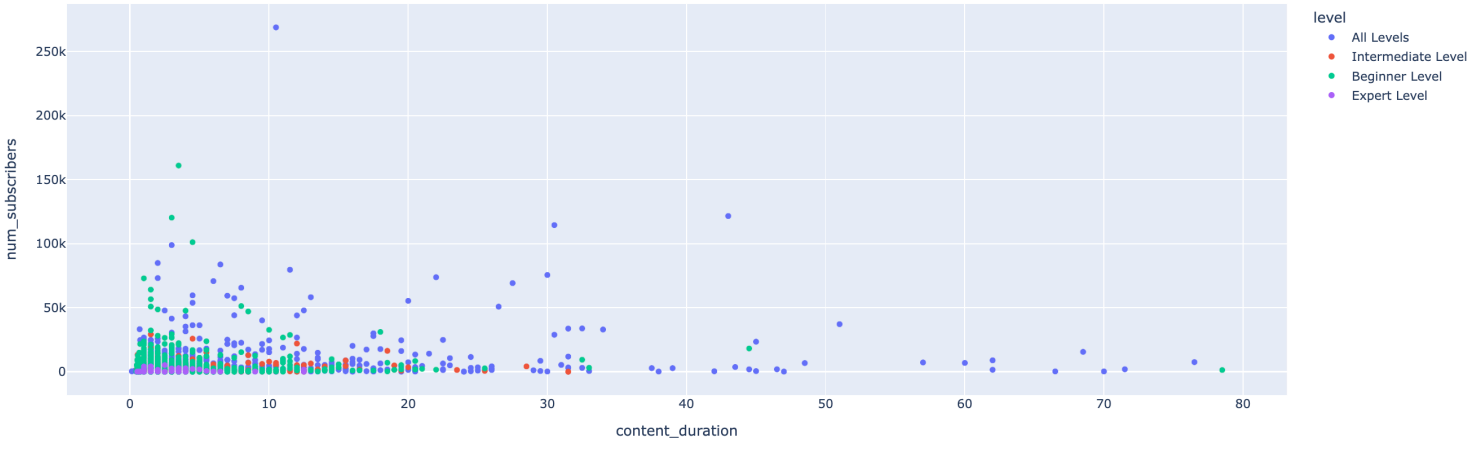
This heat map represents the correlation between the numerical columns in the dataset.

Here we can see that the maximum positive correlation is between the columns price and is_paid followed by price and content_duration

And we can also see that the maximum negative correlation is between the columns num_subscribers and is_paid followed by num_subscribers and year

Scatter plot using plotly.express library:

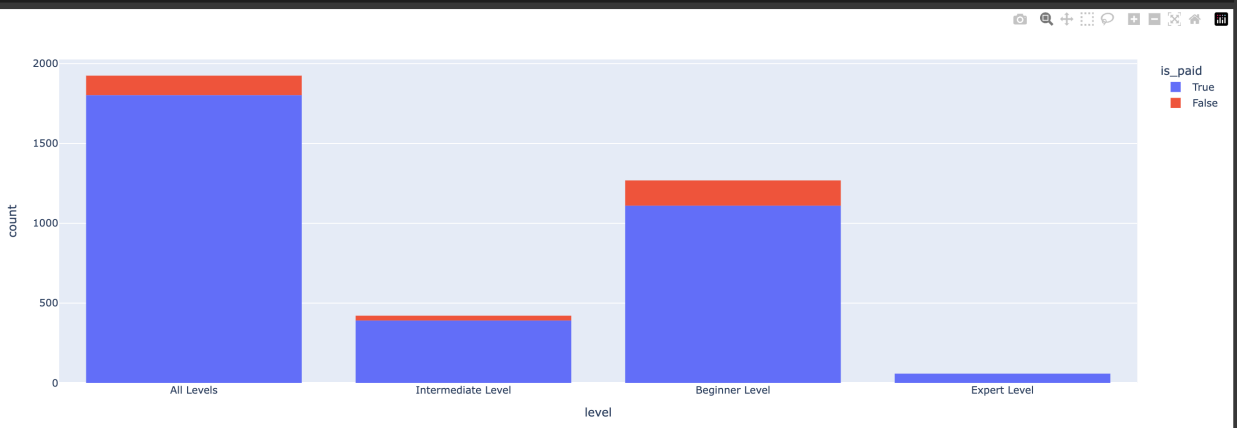
```
import plotly.express as px

fig = px.scatter(df, x="content_duration", y="num_subscribers", color="level")
fig.show()
```



From this graph we can see the distribution of num_subscribers is more between the range of 0 to 10 and 10 to 20.

This means that most of the subscribers prefer courses with less duration and also we can see that most of the courses offered between that range are beginner level courses thus we can also say that most of the people prefer beginner level courses.

```
fig1 = px.histogram(df, x="level", nbins=30,color='is_paid')
fig1.show()
```

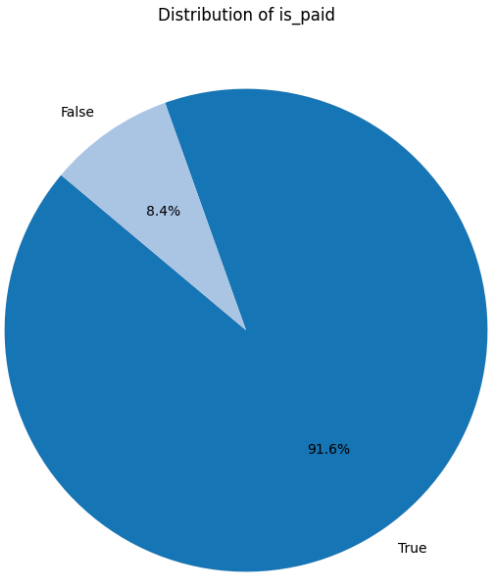


In this histogram we can clearly see that out of the courses offered the course most of the courses are suitable for all level of users and then comes the course designed solely for beginners.

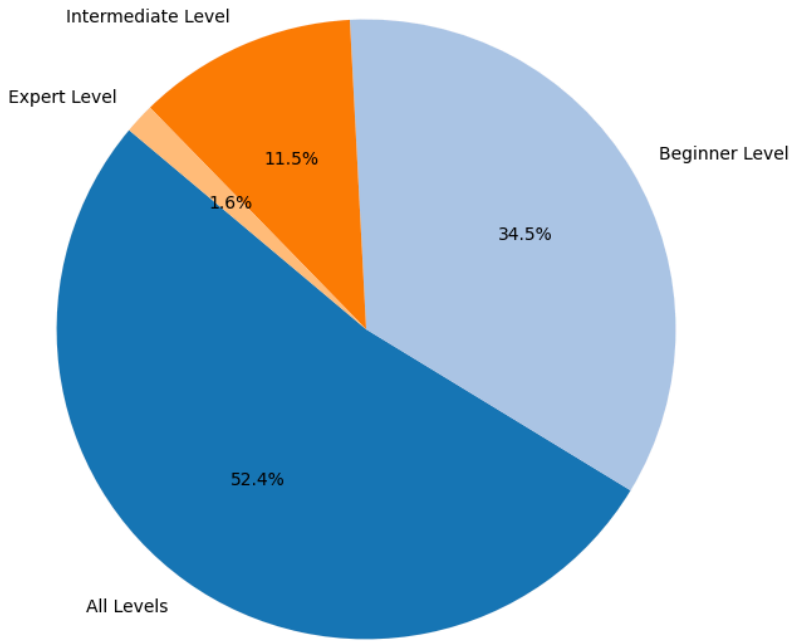
The graph also depicts the amount of paid and unpaid course proportions among each level, There are no free or unpaid courses in expert level.

Use of pie chart for analysis:

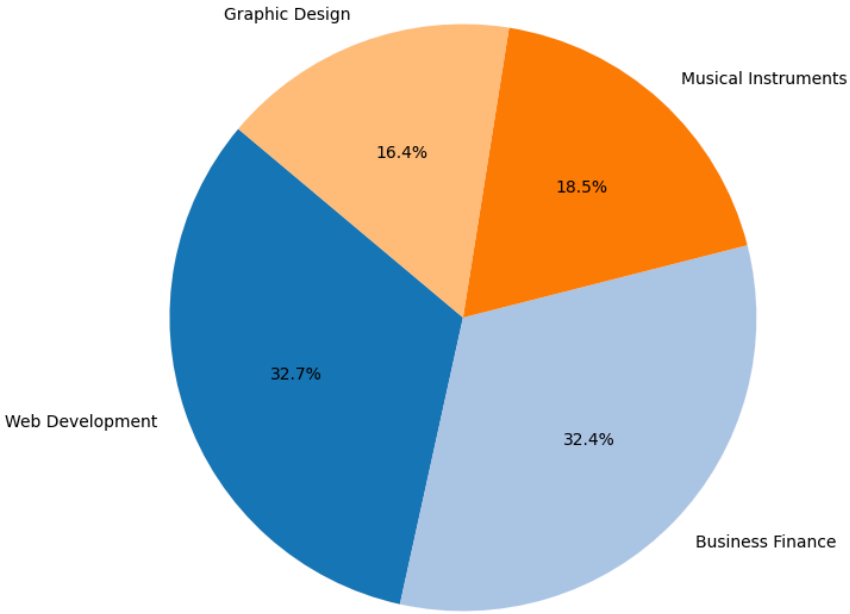
```
lst = ['is_paid', 'level', 'subject', 'year']
for i in lst:
    plot = df[i].value_counts()
    plt.figure(figsize=(8, 8))
    plt.pie(
        plot,
        labels=plot.index,
        autopct='%1.1f%%',
        startangle=140,
        colors=plt.cm.tab20.colors
    )
    plt.title(f'Distribution of {i}')
    plt.show()
    print()
```

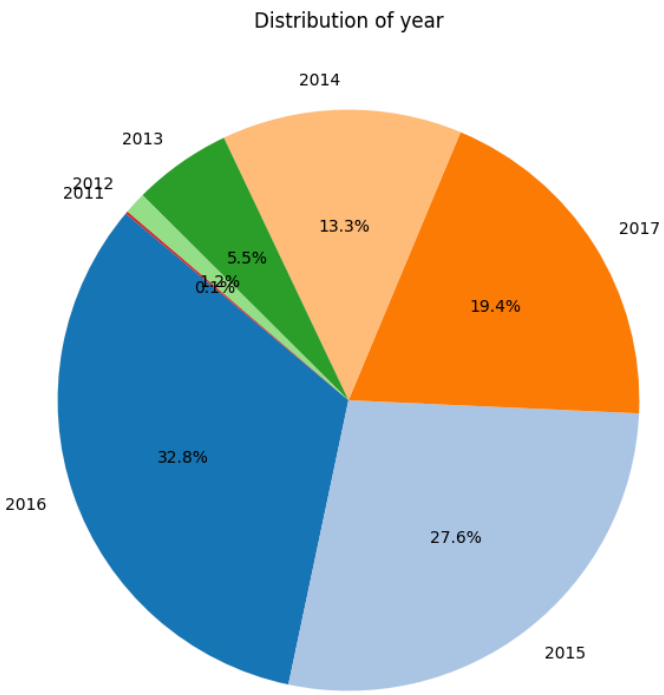


Distribution of level



Distribution of subject





From the pie charts,

- 91.6% of the courses are paid.
- Out of all courses, most of the courses offered are suitable for all level of users.
- Business Finance and Web development are the widely provided subjects.
- Most of the courses where published in the year 2016, followed by the 2015.

PYTHON LIBRARIES USED:

Pandas:

Pandas is a powerful Python library used for data manipulation and analysis. It provides data structures like `DataFrame` and `Series` to handle structured data efficiently. Pandas is widely used for tasks like cleaning, filtering, aggregating, and transforming data, making it a cornerstone for data science workflows.

Matplotlib:

Matplotlib is a versatile plotting library for creating static, interactive, and animated visualizations in Python. It provides extensive tools for crafting bar plots, scatter plots, histograms, and more. Though highly customizable, it requires more manual configuration compared to modern libraries like Seaborn or Plotly.

Seaborn

Seaborn is a Python data visualization library built on top of Matplotlib. It simplifies the creation of aesthetically pleasing statistical graphics, such as heatmaps, violin plots, and pair plots. With built-in themes and enhanced interactivity, Seaborn is ideal for exploratory data analysis.

Plotly Express

Plotly Express is a high-level module of Plotly, designed for creating interactive visualizations with minimal code. It supports a wide range of chart types, including scatter plots, pie charts, and choropleths. Its interactivity and user-friendly API make it suitable for dashboards and web applications.

CONCLUSION

The exploratory data analysis (EDA) of the Udemy online education dataset reveals valuable insights into course offerings, popularity, and trends. The dataset comprises courses across various subjects, with notable variations in their distribution. Subjects like Web Development and Business dominate in terms of the number of courses, while others are relatively underrepresented. A significant proportion of the courses are free, but paid courses exhibit a diverse pricing structure. Subscriber counts are highly skewed, with a few courses amassing an exceptionally high number of subscribers, indicating the presence of outliers. Similarly, num_reviews, num_lectures, and content_duration show wide variability, with a handful of courses having significantly higher values than the average. Yearly

trends suggest changes in subject popularity and course production over time. Visualizations like bar charts and pie charts helped uncover these patterns. Overall, this analysis highlights key trends in course offerings and user preferences, setting the stage for further investigations into factors like pricing strategies, content length optimization, and subject-specific demand.