In [1]:
```
import pandas as pd
df=pd.read_csv(r"D:\Documents\python\projects\Udemy.csv", index_col=0)
df.head()
```

Out[1]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
288942	#1 Piano Hand Coordination: Play 10th Ballad i...	True	35	3137	18	68	All Levels	1.5 hours	2014-09-18T05:07:05Z	Musical Instruments
1170074	#10 Hand Coordination - Transfer Chord Ballad ...	True	75	1593	1	41	Intermediate Level	1 hour	2017-04-12T19:06:34Z	Musical Instruments
1193886	#12 Hand Coordination: Let your Hands dance wi...	True	75	482	1	47	Intermediate Level	1.5 hours	2017-04-26T18:34:57Z	Musical Instruments
1116700	#4 Piano Hand Coordination: Fun Piano Runs in ...	True	75	850	3	43	Intermediate Level	1 hour	2017-02-21T23:48:18Z	Musical Instruments
1120410	#5 Piano Hand Coordination: Piano Runs in 2 ...	True	75	940	3	32	Intermediate Level	37 mins	2017-02-21T23:44:49Z	Musical Instruments

In [23]:
```
df.columns
df.num_lectures.max()#maximum number of lecturers
df.num_lectures.min()
df.num_lectures.describe()
```

Out[23]:
```
count    3682.000000
mean       40.065182
std        50.373299
min         0.000000
25%        15.000000
50%        25.000000
75%        45.000000
max       779.000000
Name: num_lectures, dtype: float64
```

The statistics provided are related to the variable num_lectures in a dataset (presumably a DataFrame) named df. Here's the breakdown:

Count: There are 3682 observations (lectures) in the dataset. Mean: The average number of lectures is approximately 40.07. Standard Deviation (Std): The variability of the number of lectures around the mean is approximately 50.37. This indicates that the number of lectures varies quite a bit across the dataset. Minimum (Min): The minimum number of lectures is 0. This suggests that there are some cases where there are no lectures at all. 25th Percentile (Q1): 25% of the observations have 15 or fewer lectures. Median (50th Percentile, Q2): 50% of the observations have 25 or fewer lectures. This is also the same as the 2nd quartile. 75th Percentile (Q3): 75% of the observations have 45 or fewer lectures. Maximum (Max): The maximum number of lectures is 779. This indicates there's quite a large range in the number of lectures, with some having significantly more than others.

In [24]:
```
print(f"The different subjects for which Udemy is offering courses is \n ",df.subject.unique())
```

```
The different subjects for which Udemy is offering courses is
  ['Musical Instruments' 'Business Finance' 'Graphic Design'
 'Web Development']
```

In [31]:
```
print(f"The subject that has the maximum number of courses is \n",df.subject.value_counts() )
```

```
The subject that has the maximum number of courses is
 subject
Web Development        1200
Business Finance       1199
Musical Instruments     680
Graphic Design          603
Name: count, dtype: int64
```

Another check through the data

The courses which are 'Free of Cost ' are

In [32]:
```
df[df.is_paid==False].head()
```

Out[32]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	Yea
286070	5 lecciones que todo guitarrista debe tomar	False	0	4452	263	14	Beginner Level	1 hour	2014-08-23 05:08:14+00:00	Musical Instruments	201
696630	7 Ways A Beginner Guitarist Can Sound Better, ...	False	0	4529	193	7	Beginner Level	36 mins	2015-12-21 18:50:50+00:00	Musical Instruments	201
955914	A beginner`s guide to fingerpicking and strumm...	False	0	3481	29	20	Beginner Level	2 hours	2016-09-13 21:51:59+00:00	Musical Instruments	201
270976	A how to guide in HTML	False	0	7318	205	8	Beginner Level	35 mins	2014-08-10 20:19:10+00:00	Web Development	201
1214144	¡Triunfar en La Bolsa de Valores No Requiere d...	False	0	338	7	6	Beginner Level	1 hour	2017-05-30 14:30:12+00:00	Business Finance	201

In [34]:
```
print(f"The number of courses which are 'Paid ' are \n", df.is_paid.sum())
df[df.is_paid==True]
```

```
The number of courses which are 'Paid ' are
 3372
```

Out[34]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
288942	#1 Piano Hand Coordination: Play 10th Ballad i...	True	35	3137	18	68	All Levels	1.5 hours	2014-09-18 05:07:05+00:00	Musical Instruments
1170074	#10 Hand Coordination - Transfer Chord Ballad ...	True	75	1593	1	41	Intermediate Level	1 hour	2017-04-12 19:06:34+00:00	Musical Instruments
1193886	#12 Hand Coordination: Let your Hands dance wi...	True	75	482	1	47	Intermediate Level	1.5 hours	2017-04-26 18:34:57+00:00	Musical Instruments
1116700	#4 Piano Hand Coordination: Fun Piano Runs in ...	True	75	850	3	43	Intermediate Level	1 hour	2017-02-21 23:48:18+00:00	Musical Instruments
1120410	#5 Piano Hand Coordination: Piano Runs in 2 ...	True	75	940	3	32	Intermediate Level	37 mins	2017-02-21 23:44:49+00:00	Musical Instruments
...	...	...	...	...	...	...	...	...	...	...
498488	Your First Successful Forex Trades - With Case...	True	200	1079	34	16	All Levels	2.5 hours	2015-05-26 20:48:48+00:00	Business Finance
328960	Your Own Site in 45 Min: The Complete Wordpres...	True	120	1566	29	36	All Levels	4 hours	2015-04-20 22:15:17+00:00	Web Development
552700	Your Second Course on Piano: Two Handed Playing	True	70	1018	12	22	Beginner Level	5 hours	2015-10-26 20:04:21+00:00	Musical Instruments
631754	Zend Framework 2: Learn the PHP framework ZF2 ...	True	40	723	130	37	All Levels	6.5 hours	2015-11-11 18:55:45+00:00	Web Development
964478	Zombie Apocalypse Photoshop Actions	True	50	12	1	15	All Levels	1.5 hours	2016-09-26 22:19:48+00:00	Graphic Design

3372 rows × 12 columns

In [35]:
```
print(f"The Top Selling Courses i.e most subscribers are \n ")

df.sort_values('num_subscribers',ascending=False).tail()
```

```
The Top Selling Courses i.e most subscribers are
```

Out[35]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	Y
1223240	Learn Pirates of the Caribbean by Ear on the P...	True	20	0	0	6	All Levels	32 mins	2017-05-22 17:14:43+00:00	Musical Instruments	2
1258666	Financial Statement Auditing Cycles	True	50	0	0	9	Intermediate Level	2 hours	2017-06-29 23:20:10+00:00	Business Finance	2
1215926	Kickstarter success in 5 easy steps	True	20	0	0	12	All Levels	31 mins	2017-05-16 14:55:28+00:00	Business Finance	2
1247992	Introduction to Project Management for Finance...	True	50	0	0	9	Beginner Level	2 hours	2017-07-03 21:40:32+00:00	Business Finance	2
1087466	Stop Creditors from Harassing you and Avoid Ba...	True	20	0	0	7	Beginner Level	37 mins	2017-02-02 16:22:37+00:00	Business Finance	2

In [37]:
```
print(f"The Least Selling Courses that is least subscribers are \n ")

df.sort_values('num_subscribers',ascending=True).tail()
```

```
The Least Selling Courses that is least subscribers are
```

Out[37]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	Ye
764164	The Complete Web Developer Course 2.0	True	200	114512	22412	304	All Levels	30.5 hours	2016-03-08 22:28:36+00:00	Web Development	20
173548	Build Your First Website in 1 Week with HTML5 ...	False	0	120291	5924	30	Beginner Level	3 hours	2014-04-08 16:21:30+00:00	Web Development	20
625204	The Web Developer Bootcamp	True	200	121584	27445	342	All Levels	43 hours	2015-11-02 21:13:27+00:00	Web Development	20
59014	Coding for Entrepreneurs Basic	False	0	161029	279	27	Beginner Level	3.5 hours	2013-06-09 15:51:55+00:00	Web Development	20
41295	Learn HTML5 Programming From Scratch	False	0	268923	8629	45	All Levels	10.5 hours	2013-02-14 07:03:41+00:00	Web Development	20

In [38]:
```
#Code wont work since there are different data types in price especially the string 'Free'. I will convert this to the number 0
df['price']=df['price'].str.replace('Free','0')
pd.to_numeric(df['price'])
print(f"The courses of Graphic Design where the price is below 100 are \n ")

df[(df.subject=='Graphic Design')&(df.price<'100')].tail()
```

```
The courses of Graphic Design where the price is below 100 are
```

Out[38]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	Year	yea
611804	Primeros Pasos con Photoshop CC	False	0	6429	654	31	Beginner Level	2.5 hours	2015-09-23 15:30:59+00:00	Graphic Design	2015	
399938	Professional Logo Design in Adobe Illustrator	False	0	44044	1563	45	All Levels	7.5 hours	2015-01-22 11:18:06+00:00	Graphic Design	2015	
611370	Quote Images for Pinterest, Facebook, & Instagram	False	0	12103	576	18	Beginner Level	1 hour	2015-09-24 19:47:45+00:00	Graphic Design	2015	
839536	Start Making Comics with Manga Studio 5 / Clip...	False	0	5301	125	77	All Levels	6.5 hours	2016-06-21 02:49:47+00:00	Graphic Design	2016	
1245392	Voxel 3D Model Creation Course	False	0	1031	9	8	All Levels	39 mins	2017-06-08 22:46:39+00:00	Graphic Design	2017	

In [49]:
```
#for i in df['price'].str.contains('Free'):
#    if df['price'].str.contains('Free') == True:
#        df.drop(df.price[i])
print(f"The Courses realted to 'Python' are listed below and they are {len(df[df.course_title.str.contains('Python')])} in number")

df[df.course_title.str.contains('Python')].head()
```

```
The Courses realted to 'Python' are listed below and they are 29 in number
```

Out[49]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
599504	Advanced Scalable Python Web Development Using...	True	120	1299	56	71	Intermediate Level	14 hours	2016-08-11 22:09:24+00:00	Web Development
47963	Coding for Entrepreneurs: Learn Python, Django...	True	195	23412	799	251	All Levels	45 hours	2013-04-08 00:46:14+00:00	Web Development
631128	Complete Python Web Course: Build 8 Python Web...	True	110	7489	941	173	All Levels	16 hours	2015-11-08 20:57:35+00:00	Web Development
186096	Core: A Web App Reference Guide for Django, Py...	True	195	2497	98	154	All Levels	26 hours	2014-05-29 00:58:43+00:00	Web Development
394832	Fun and creative web engineering with Python a...	False	0	10917	319	25	All Levels	2 hours	2015-06-09 19:51:50+00:00	Web Development

In [77]:
```
print(f"The Courses realted to 'Java' are listed below and they are {len(df[df.course_title.str.contains('Java')])} in number")

df[df.course_title.str.contains('Java')].tail()
```

```
The Courses realted to 'Java' are listed below and they are 149 in number
```

Out[77]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	Y
1270392	The Complete JavaScript, HTML and CSS Tutorial...	True	35	244	44	21	Beginner Level	3 hours	2017-06-29 16:13:18+00:00	Web Development	2
895096	Ultimate JavaScript Objects	True	20	1525	43	65	All Levels	2.5 hours	2016-07-07 21:00:57+00:00	Web Development	2
909836	Ultimate JavaScript Strings	False	0	3142	117	20	Beginner Level	37 mins	2016-07-29 12:53:13+00:00	Web Development	2
711592	Using Modern JavaScript Today	True	50	1658	185	68	Intermediate Level	16.5 hours	2016-01-01 18:34:53+00:00	Web Development	2
900434	VueJS V1 Introduction to VueJS JavaScript Fram...	True	200	3632	28	31	Beginner Level	2 hours	2016-07-26 16:53:56+00:00	Web Development	2

In [79]:
```
df.dtypes
#convert published_timestamp     to datetime format
df['published_timestamp']=pd.to_datetime(df.published_timestamp)
df.dtypes
# Create a new coliumn year from published_timestamp   column
df['Year']=df['published_timestamp'].dt.year
df.head(5)
print(f"The courses that published in 2015 are {len(df[df.Year==2015])} as follows \n")


df[df.Year==2015].head()
```

```
The courses that published in 2015 are 1014 as follows
```

Out[79]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	Ye
591880	1 - Concepts of Statistics For Beginners Step ...	True	200	273	4	15	Beginner Level	31 mins	2015-08-30 22:48:34+00:00	Business Finance	20
302450	10 Numbers Every Business Owner Should Know	True	20	13	1	9	All Levels	1 hour	2015-03-08 19:11:24+00:00	Business Finance	20
384928	101 Blues riffs - learn how the harmonica supe...	True	200	1350	65	55	Intermediate Level	6.5 hours	2015-01-04 21:14:31+00:00	Musical Instruments	20
550842	16 Guitar Chords to Jam With - (Beginner - Int...	True	20	1224	19	20	Beginner Level	1 hour	2015-07-10 19:53:56+00:00	Musical Instruments	20
486240	2 Easy Steps To Investment And Avoiding Traps	True	20	828	1	20	All Levels	1 hour	2015-04-27 23:18:59+00:00	Business Finance	20

In [80]:
```
df.level.unique()
```

Out[80]:
```
array(['All Levels', 'Intermediate Level', 'Beginner Level',
       'Expert Level'], dtype=object)
```

In [81]:
```
print(f"The max No of Subscribers for each Level of Courses is \n {df.groupby('level')['num_subscribers'].max()}")
```

```
The max No of Subscribers for each Level of Courses is
 level
All Levels            268923
Beginner Level        161029
Expert Level            5172
Intermediate Level     29167
Name: num_subscribers, dtype: int64
```

Qualitative Analysis Questions:

Content Popularity: Which course subjects are the most popular among subscribers? Provide a visualization illustrating the distribution of subscribers across different subjects.

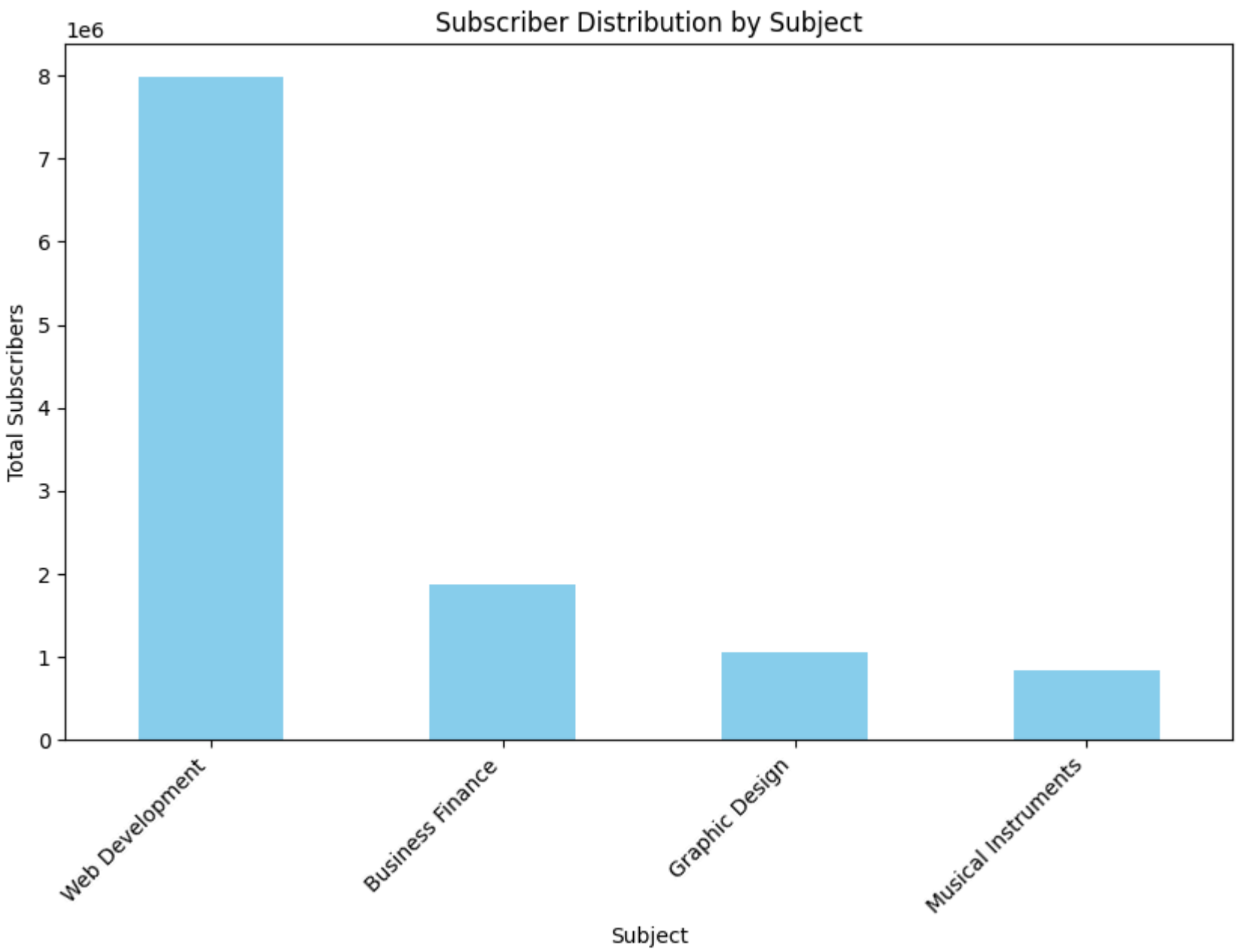
Level Distribution: Can you analyze the distribution of course levels (beginner, intermediate, advanced)? How does the popularity of each level vary across different subjects?

Review Sentiment Analysis: Could you perform sentiment analysis on course reviews? Are there any patterns or trends in sentiment across different course subjects or levels?

Course Pricing Strategy: Analyze the distribution of course prices. Is there a correlation between course pricing and the number of subscribers or reviews?

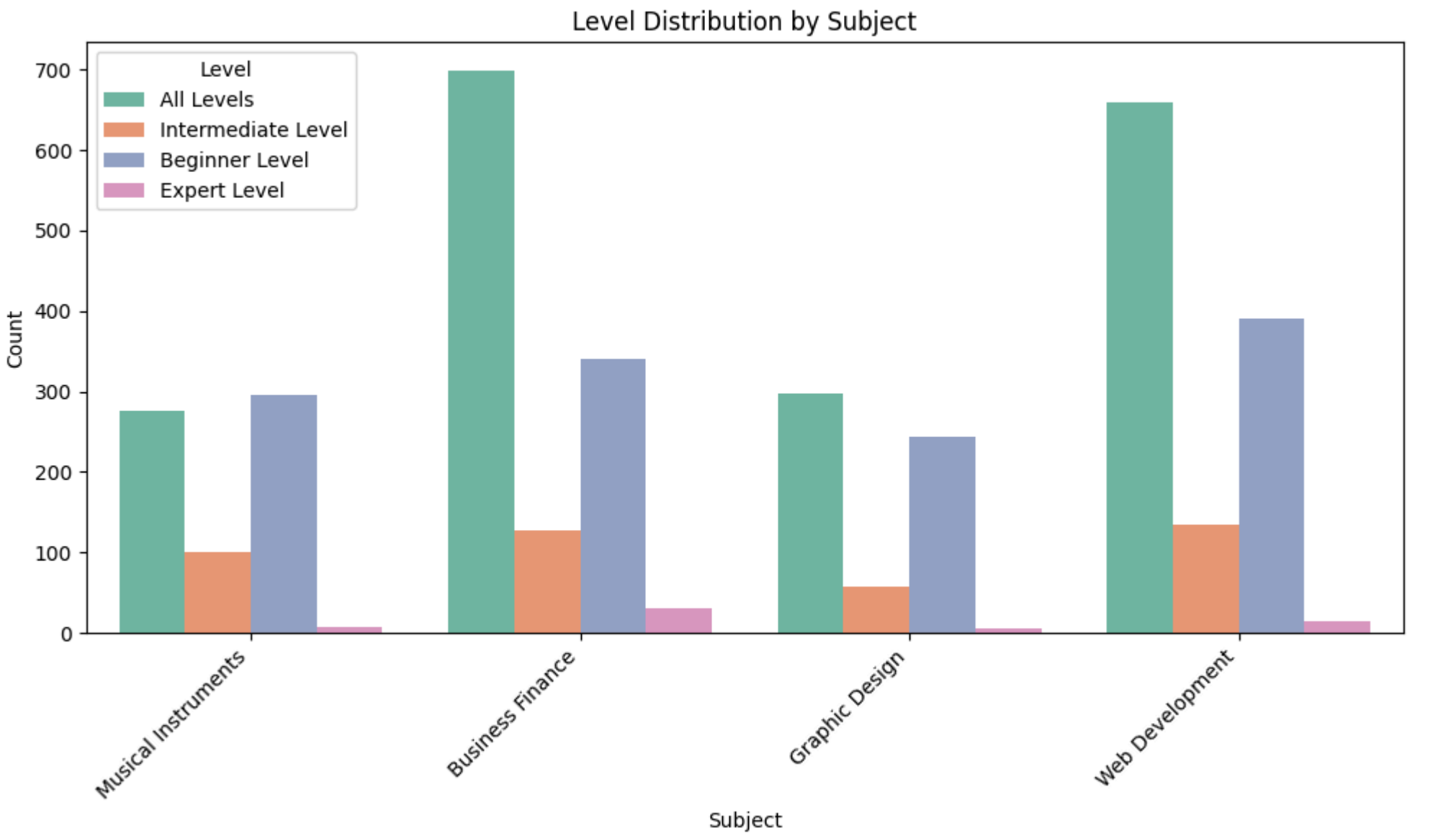
Duration vs. Engagement: Is there any relationship between the duration of a course and its engagement (measured by the number of subscribers or reviews)? Create visualizations to illustrate any trends or patterns.

In [14]:
```python
# Which course subjects are the most popular among subscribers? Provide a visualization illustrating the distribution of subscribers across
import matplotlib.pyplot as plt

# Grouping data by subject and calculating total subscribers
subject_subscribers = df.groupby('subject')['num_subscribers'].sum()

# Creating a bar plot to visualize subscriber distribution by subject
plt.figure(figsize=(10, 6))
subject_subscribers.sort_values(ascending=False).plot(kind='bar', color='skyblue')
plt.title('Subscriber Distribution by Subject')
plt.xlabel('Subject')
plt.ylabel('Total Subscribers')
plt.xticks(rotation=45, ha='right')
#plt.tight_layout()
plt.show()
```

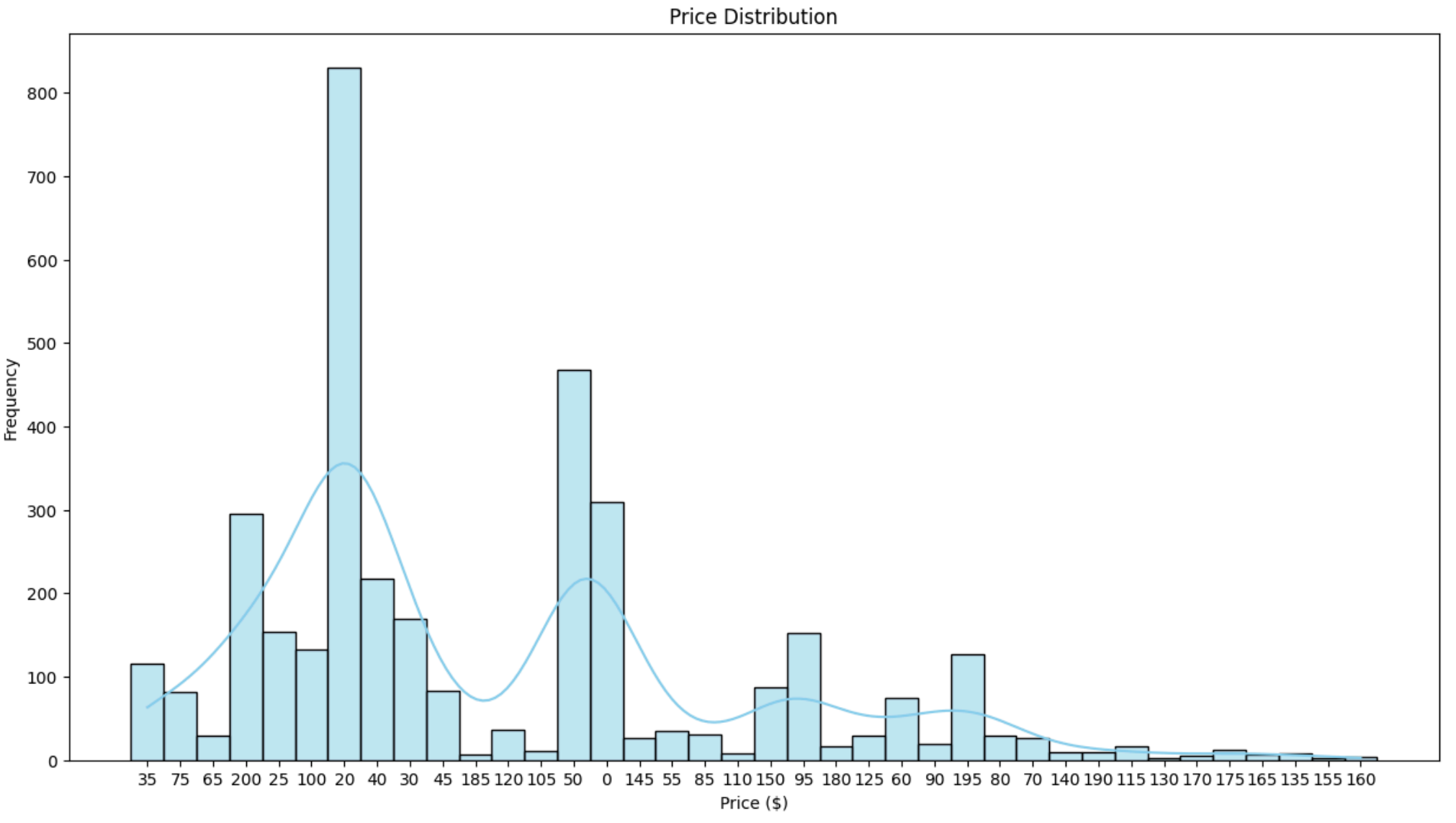


In [15]:
```python
# Can you analyze the distribution of course levels (beginner, intermediate, advanced)? How does the popularity of each level vary across di
import seaborn as sns

# Creating a countplot to visualize level distribution by subject
plt.figure(figsize=(10, 6))
sns.countplot(data=df, x='subject', hue='level', palette='Set2')
plt.title('Level Distribution by Subject')
plt.xlabel('Subject')
```

```
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.legend(title='Level')
plt.tight_layout()
plt.show()
```



In [16]:
```
#
#Course Pricing Strategy: Analyze the distribution of course prices. Is there a correlation between course pricing and the number of subscr

# Visualizing price distribution
plt.figure(figsize=(15, 8))
sns.histplot(df['price'], bins=10, kde=True, color='skyblue')
plt.title('Price Distribution')
plt.xlabel('Price ($)')
plt.ylabel('Frequency')
plt.show()
```



In [17]:
```
# Analyzing correlation between price and other variables (e.g., num_subscribers, num_reviews)
import numpy as np
```

```
# Compute the correlation coefficients between price and other variables
correlation_price = df['price'].corr(df['num_subscribers'])
correlation_reviews = df['price'].corr(df['num_reviews'])

print("Correlation between price and number of subscribers:", correlation_price)
print("Correlation between price and number of reviews:", correlation_reviews)
```

```
Correlation between price and number of subscribers: 0.05093901742403312
Correlation between price and number of reviews: 0.11377754451042968
```

Correlation between Price and Number of Subscribers:

Correlation Coefficient: 0.0509 Interpretation: There is a very weak positive correlation between the price of the course and the number of subscribers. This suggests that as the price increases, there's a slight tendency for the number of subscribers to increase, but the relationship is very weak.

Correlation between Price and Number of Reviews:

Correlation Coefficient: 0.1138 Interpretation: There is a weak positive correlation between the price of the course and the number of reviews. This indicates that as the price increases, there's a slight tendency for the number of reviews to increase, but again, the relationship is weak.

In [18]:
```
#**Kolmogorov-Smirnov Test:**
#
#Duration vs. Engagement: Is there any relationship between the duration of a course and its engagement (measured by the number of subscribe

# Compute the correlation coefficients between num_lectures and other variables
correlation_lectures_subscribers = df['num_lectures'].corr(df['num_subscribers'])
correlation_lectures_reviews = df['num_lectures'].corr(df['num_reviews'])

print("Correlation between num_lectures and num_subscribers:", correlation_lectures_subscribers)
print("Correlation between num_lectures and num_reviews:", correlation_lectures_reviews)
```

```
Correlation between num_lectures and num_subscribers: 0.15792877640002856
Correlation between num_lectures and num_reviews: 0.24308286692371922
```

Correlation between Number of Lectures and Number of Subscribers:

Correlation Coefficient: 0.1579 Interpretation: There is a moderate positive correlation between the number of lectures in a course and the number of subscribers. This suggests that courses with more lectures tend to attract more subscribers, indicating a moderate relationship between these variables.

Correlation between Number of Lectures and Number of Reviews:

Correlation Coefficient: 0.2431 Interpretation: There is a moderate positive correlation between the number of lectures in a course and the number of reviews. This indicates that courses with more lectures tend to receive more reviews, suggesting a moderate relationship between these variables.

Quantitative Analysis Questions:

Subscriber Growth Over Time: How has the number of subscribers evolved over time? Create a time series visualization showing the growth trends.

Price Sensitivity Analysis: Can you analyze price sensitivity among subscribers? For example, how does price affect the number of subscribers or reviews for courses in different subjects?

Lecture Engagement: Analyze the distribution of the number of lectures across different courses. Is there a relationship between the number of lectures and course popularity or engagement?

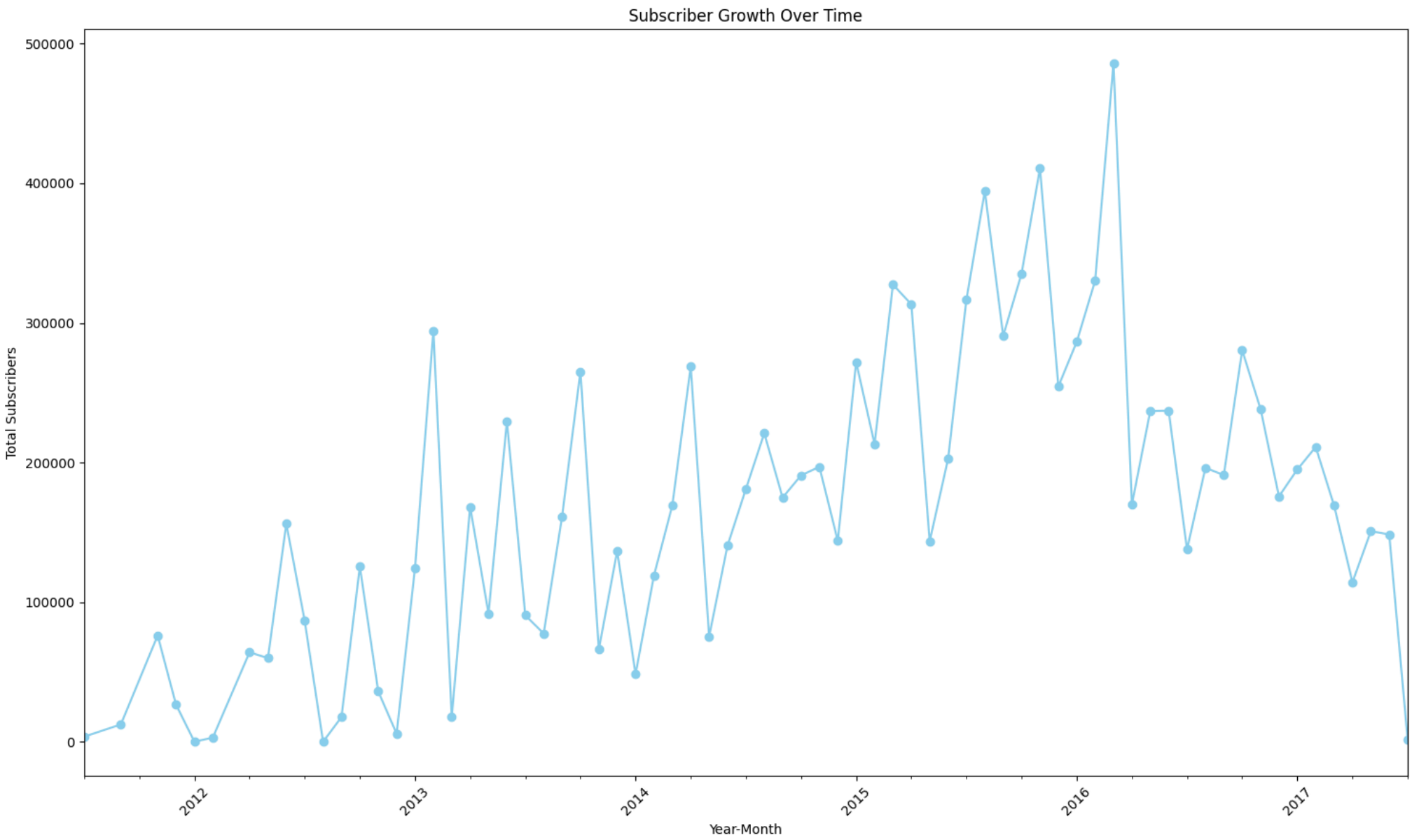
Subject Comparison: Perform a comparative analysis of course subjects. Which subjects have the highest average number of subscribers or reviews? Visualize the comparison to highlight any significant differences.

Correlation Analysis: Are there any correlations between variables such as the number of lectures, course duration, and the number of subscribers or reviews? Create correlation matrices and visualizations to explore these relationships.

Subscriber Growth Over Time: How has the number of subscribers evolved over time? Create a time series visualization showing the growth trends.

In [19]:
```
# Convert published_timestamp to datetime if it's not already
df['published_timestamp'] = pd.to_datetime(df['published_timestamp'])

# Extract year and month from published_timestamp
df['year_month'] = df['published_timestamp'].dt.to_period('M')

# Grouping data by year_month and calculating total subscribers
subscriber_growth = df.groupby('year_month')['num_subscribers'].sum()

# Creating a line plot to visualize subscriber growth over time
plt.figure(figsize=(15, 9))
subscriber_growth.plot(marker='o', color='skyblue')
plt.title('Subscriber Growth Over Time')
plt.xlabel('Year-Month')
plt.ylabel('Total Subscribers')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
C:\Users\Harold\AppData\Local\Temp\ipykernel_1872\1953189972.py:5: UserWarning: Converting to PeriodArray/Index representation will drop timezone information.
  df['year_month'] = df['published_timestamp'].dt.to_period('M')
```

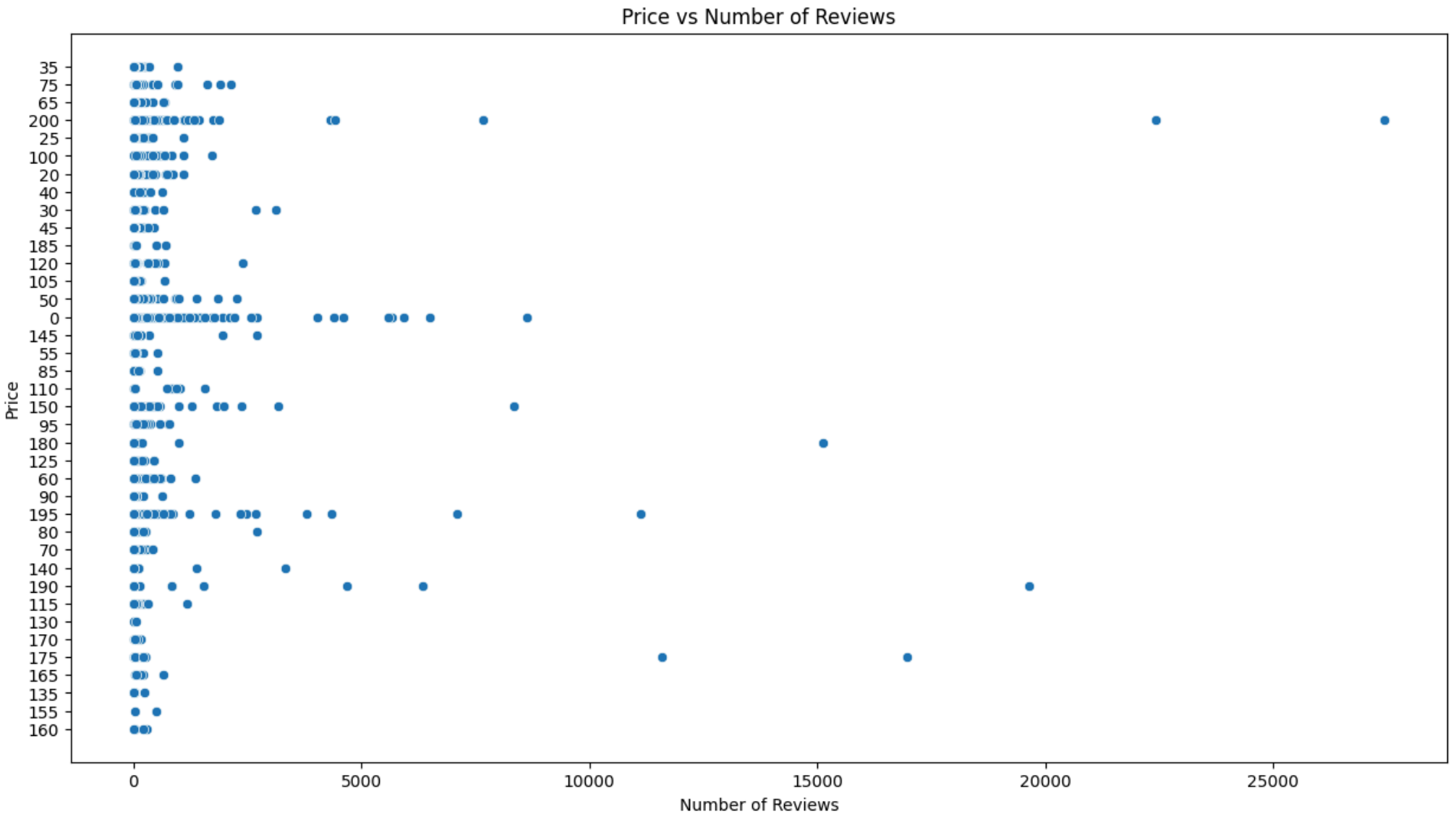


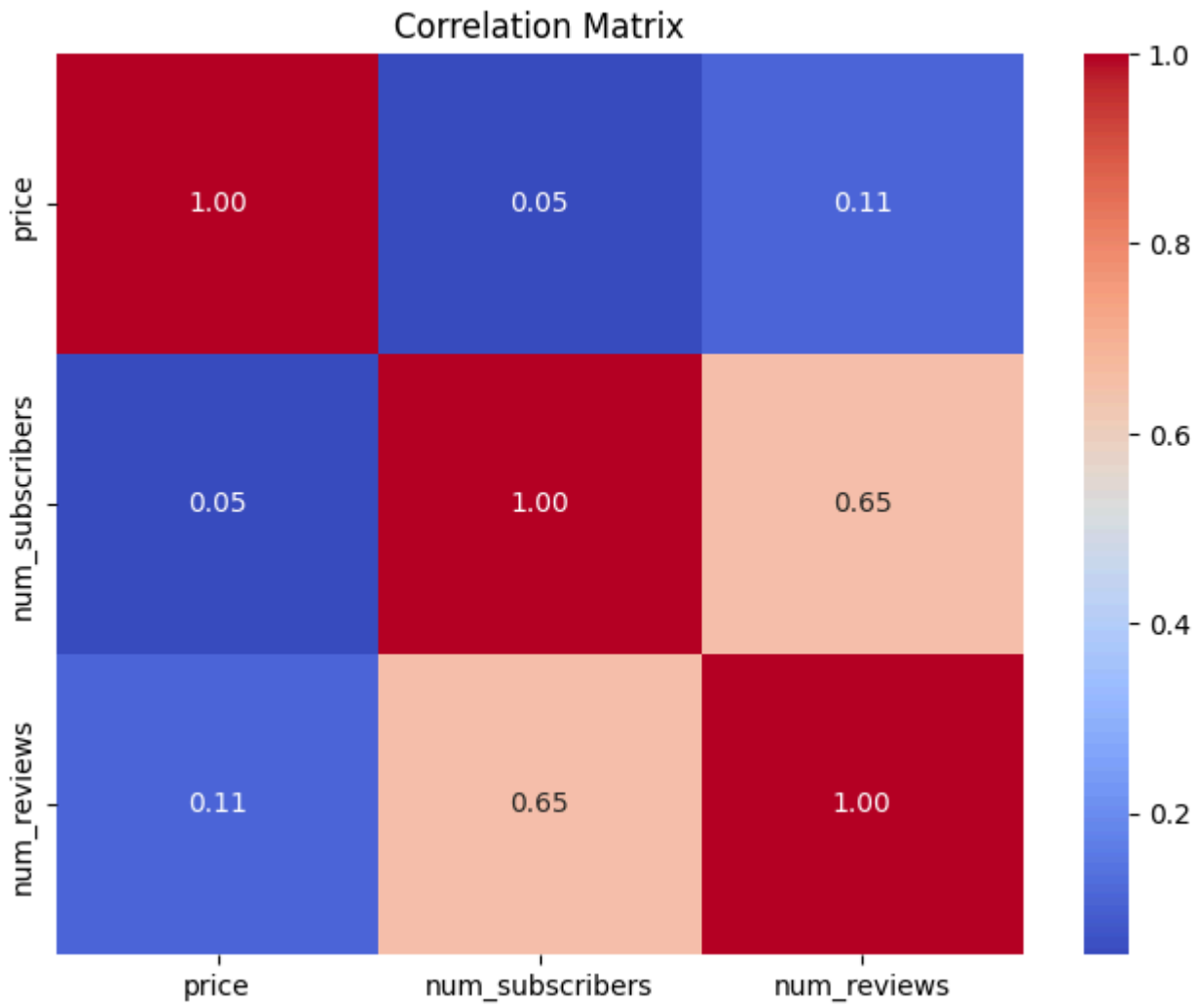
Analyzing correlation between price and other variables (e.g., num_subscribers, num_reviews).

Visualization can include scatter plots, regression analysis, or correlation matrices

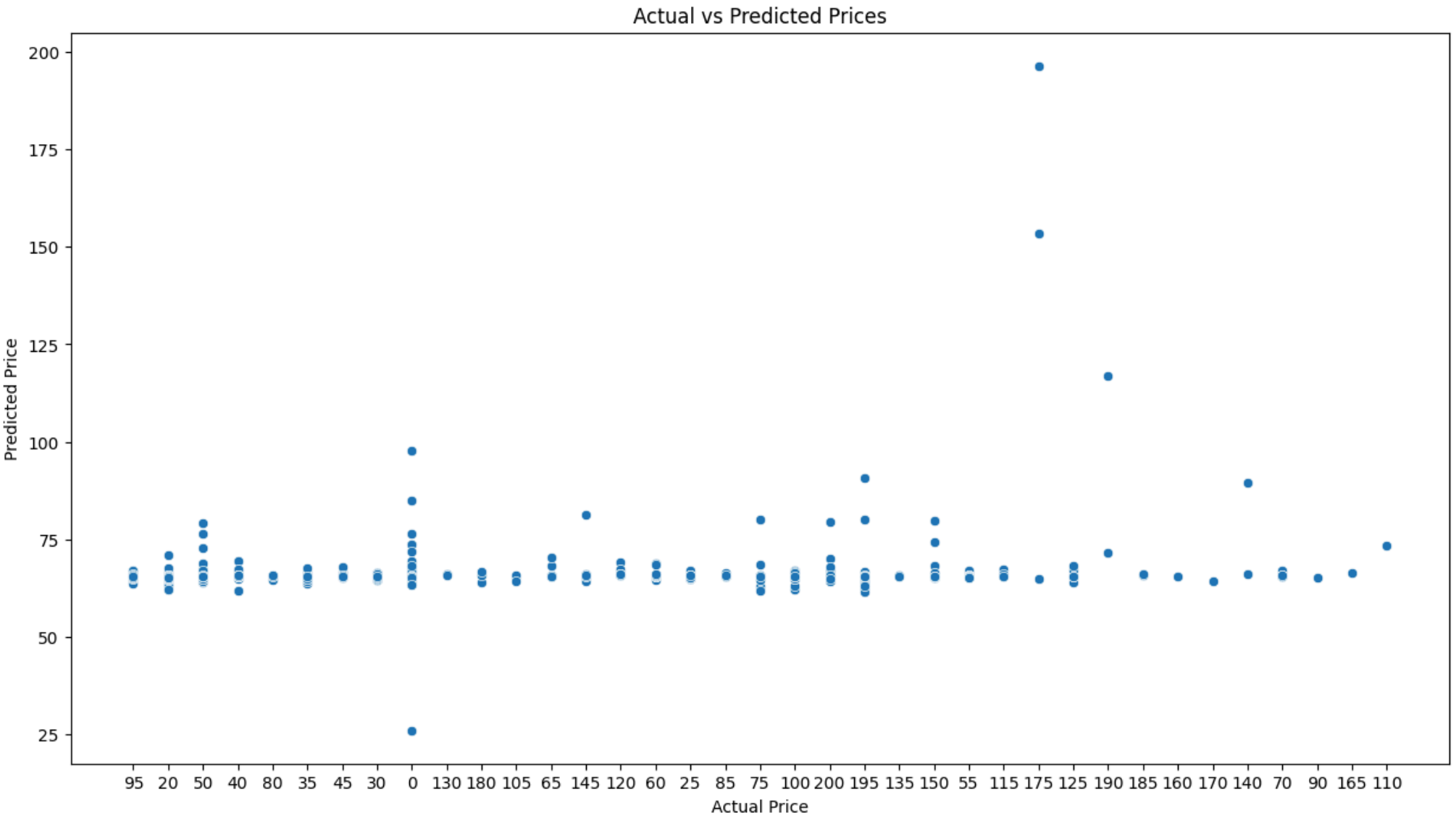
In [20]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error


# Create a scatter plot to visualize the relationship between price and num_subscribers
plt.figure(figsize=(15, 8))
sns.scatterplot(x='num_subscribers', y='price', data=df)
plt.title('Price vs Number of Subscribers')
plt.xlabel('Number of Subscribers')
plt.ylabel('Price')
plt.show()

# Create a scatter plot to visualize the relationship between price and num_reviews
plt.figure(figsize=(15, 8))
sns.scatterplot(x='num_reviews', y='price', data=df)
plt.title('Price vs Number of Reviews')
plt.xlabel('Number of Reviews')
plt.ylabel('Price')
plt.show()

# Compute the correlation matrix
correlation_matrix = df[['price', 'num_subscribers', 'num_reviews']].corr()

# Visualize the correlation matrix using a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()

# Split the data into training and testing sets
X = df[['num_subscribers', 'num_reviews']]
y = df['price']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train a linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = model.predict(X_test)
```

```python
# Compute the mean squared error
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)

# Plot the actual vs predicted prices
plt.figure(figsize=(15, 8))
sns.scatterplot(x=y_test, y=y_pred)
plt.xlabel('Actual Price')
plt.ylabel('Predicted Price')
plt.title('Actual vs Predicted Prices')
plt.show()
```






```
Mean Squared Error: 3504.5947063646527
```



Interpretation of the Correlation Matrix

Correlation Coefficient: 0.11

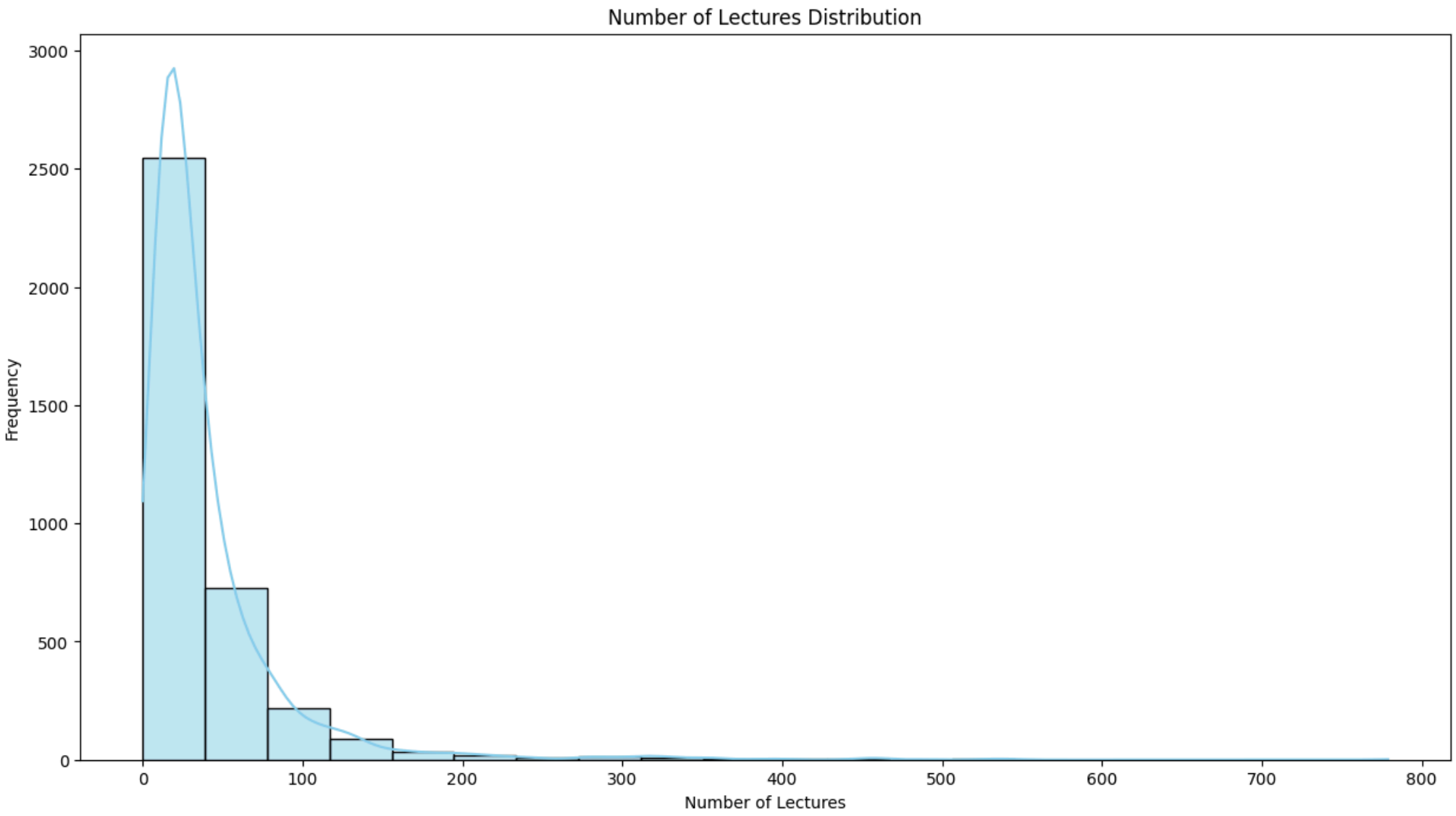
Interpretation: There is a weak positive correlation between the price of a course and the number of reviews. This indicates that as the price increases, there's a slight tendency for the number of reviews to increase, but again, the relationship is weak.

Correlation Coefficient: 0.65 Interpretation: There is a moderate positive correlation between the number of lectures in a course and the number of reviews. This indicates that courses with more lectures tend to receive more reviews, suggesting a moderate relationship between these variables.

Lecture Engagement: Analyze the distribution of the number of lectures across different courses. Is there a relationship between the number of lectures and course popularity or engagement?

In [44]:
```
# Visualizing distribution of num_lectures
plt.figure(figsize=(15, 8))
sns.histplot(df['num_lectures'], bins=20, kde=True, color='skyblue')
plt.title('Number of Lectures Distribution')
plt.xlabel('Number of Lectures')
plt.ylabel('Frequency')
plt.show()
```

```
# Analyzing correlation between num_lectures and other variables (e.g., num_subscribers, num_reviews)

import pandas as pd
import numpy as np


# Compute the correlation coefficients between num_lectures and other variables
correlation_lectures_subscribers = df['num_lectures'].corr(df['num_subscribers'])
correlation_lectures_reviews = df['num_lectures'].corr(df['num_reviews'])

print("Correlation between num_lectures and num_subscribers:", correlation_lectures_subscribers)
print("Correlation between num_lectures and num_reviews:", correlation_lectures_reviews)
```

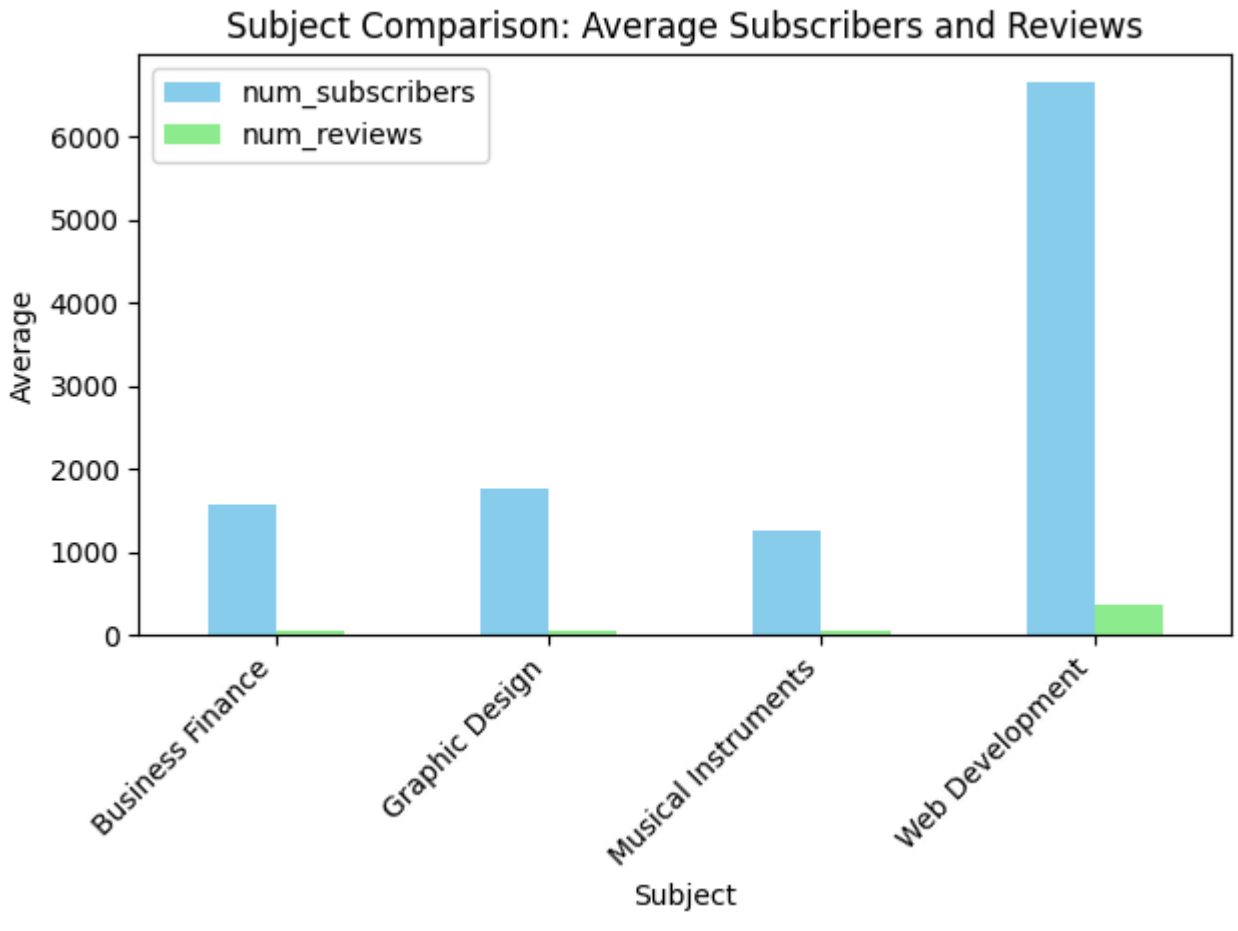



```
Correlation between num_lectures and num_subscribers: 0.15792877640002856
Correlation between num_lectures and num_reviews: 0.24308286692371922
```

Subject Comparison: Perform a comparative analysis of course subjects. Which subjects have the highest average number of subscribers or reviews? Visualize the comparison to highlight any significant differences.

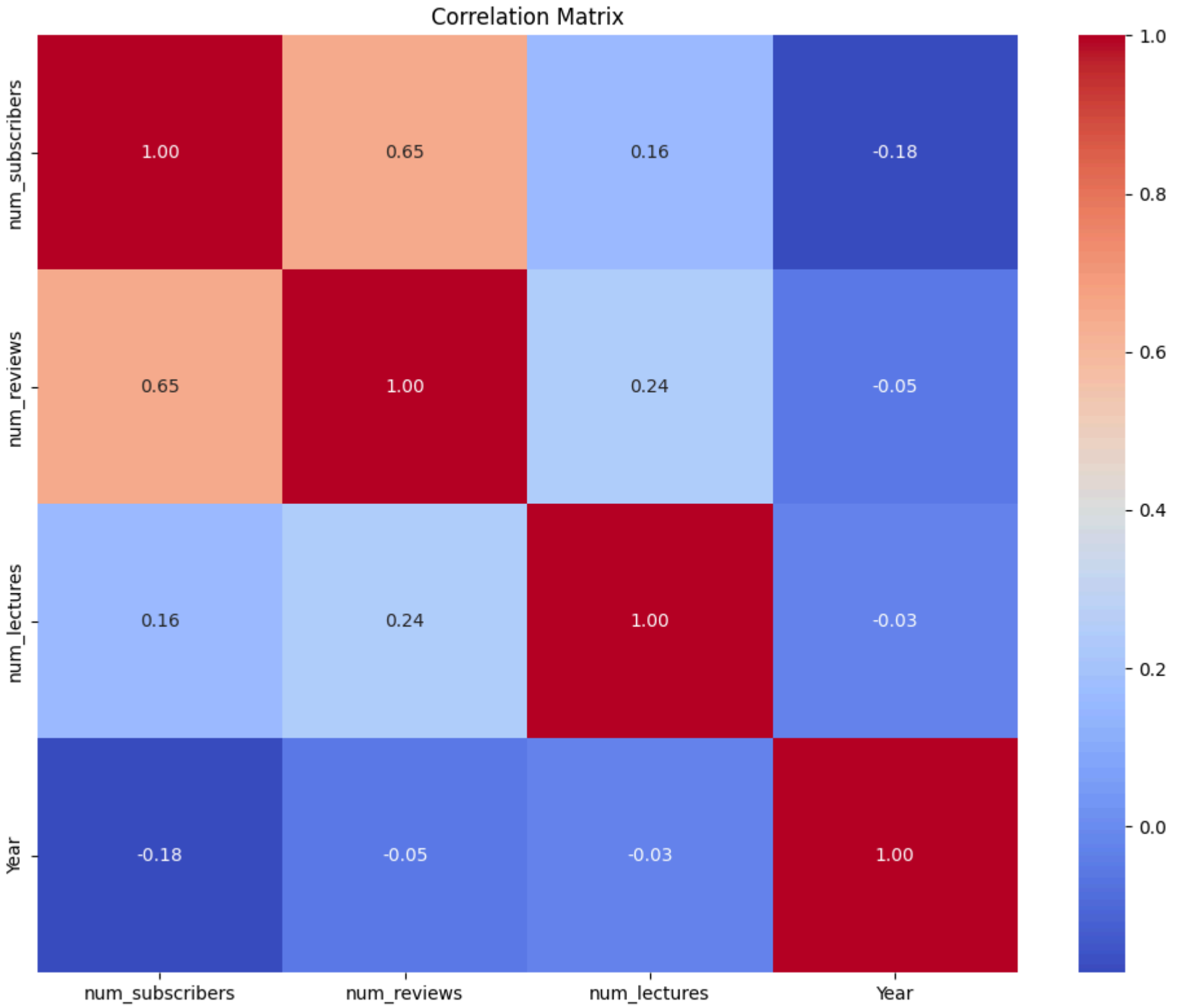
In [45]:
```
# Grouping data by subject and calculating average number of subscribers and reviews
subject_stats = df.groupby('subject').agg({'num_subscribers': 'mean', 'num_reviews': 'mean'})

# Creating bar plots to compare average subscribers and reviews by subject
plt.figure(figsize=(10, 6))
subject_stats.plot(kind='bar', color=['skyblue', 'lightgreen'])
plt.title('Subject Comparison: Average Subscribers and Reviews')
plt.xlabel('Subject')
plt.ylabel('Average')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



Correlation Analysis: Are there any correlations between variables such as the number of lectures, course duration, and the number of subscribers or reviews? Create correlation matrices and visualizations to explore these relationships.

In [21]:
```
# Select only numerical columns
numeric_df = df.select_dtypes(include=[np.number])

# Calculating correlation matrix
correlation_matrix = numeric_df.corr()

# Visualizing correlation matrix using heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.tight_layout()
plt.show()
```



Theere is little-to-no correlation between the year of a course upload and the number of subscribers, -0.18, the number of reviews, -0.05, and the number of lectures which has a correlation coefficient of -0.03.

Interpretation of the Correlation Matrix

Comparing Number of Reviews and the Number of Lectures

Correlation Coefficient: 0.24

Interpretation: There is a moderate positive correlation between the number of lectures of a course and the number of reviews. This indicates that courses with more lectures tend to receive more reviews.

Correlation Coefficient: 0.16 Interpretation: There is a weak positive correlation between the number of lectures in a course and the number of subscribers. This indicates that as the number of lectures increases, there's a slight tendency for the number of reviews to increase.

In [49]:
```
df.tail()
```

Out[49]:

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	Y
288942	#1 Piano Hand Coordination: Play 10th Ballad i...	True	35	3137	18	68	All Levels	1.5 hours	2014-09-18 05:07:05+00:00	Musical Instruments	2
1170074	#10 Hand Coordination - Transfer Chord Ballad ...	True	75	1593	1	41	Intermediate Level	1 hour	2017-04-12 19:06:34+00:00	Musical Instruments	2
1193886	#12 Hand Coordination: Let your Hands dance wi...	True	75	482	1	47	Intermediate Level	1.5 hours	2017-04-26 18:34:57+00:00	Musical Instruments	2
1116700	#4 Piano Hand Coordination: Fun Piano Runs in ...	True	75	850	3	43	Intermediate Level	1 hour	2017-02-21 23:48:18+00:00	Musical Instruments	2
1120410	#5 Piano Hand Coordination: Piano Runs in 2 ...	True	75	940	3	32	Intermediate Level	37 mins	2017-02-21 23:44:49+00:00	Musical Instruments	2

In [ ]: