Building a Course Recommend System

NTI
National Telecommunication Institute

Team Members :

- Safinaz Ahmed
- Mina Waheed
- Kerolos Adel

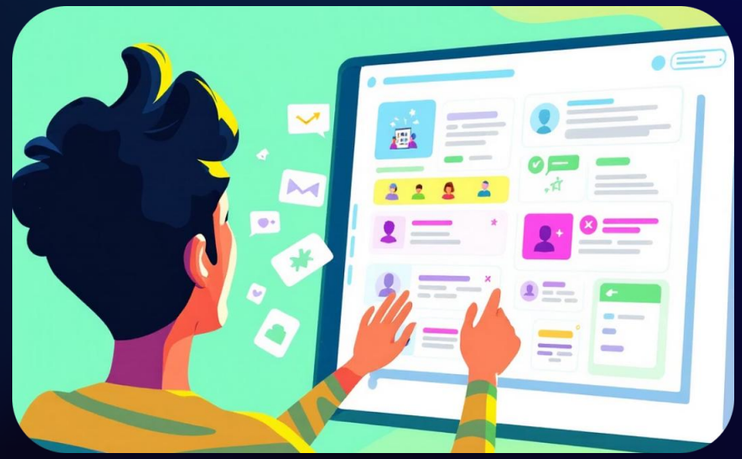
💡 Project Overview

Main Idea: Course Recommender System

Build a **Course Recommender System** that helps users discover online courses based on their interests by analyzing course content and user behavior.

Goal: Relevant & High-Quality Courses

Help users find **relevant** and **high-quality** courses easily. We use **content-based filtering** based on features like course title, subject, level and price.



Why This Matters: Personalized Discovery

Online learners face **too many choices**. Our system makes discovery **faster, smarter, and personalized**, streamlining the learning journey.

Agenda

Today, we'll explore the journey of building a personalized course recommender system.

Preprocessing & Visualizations

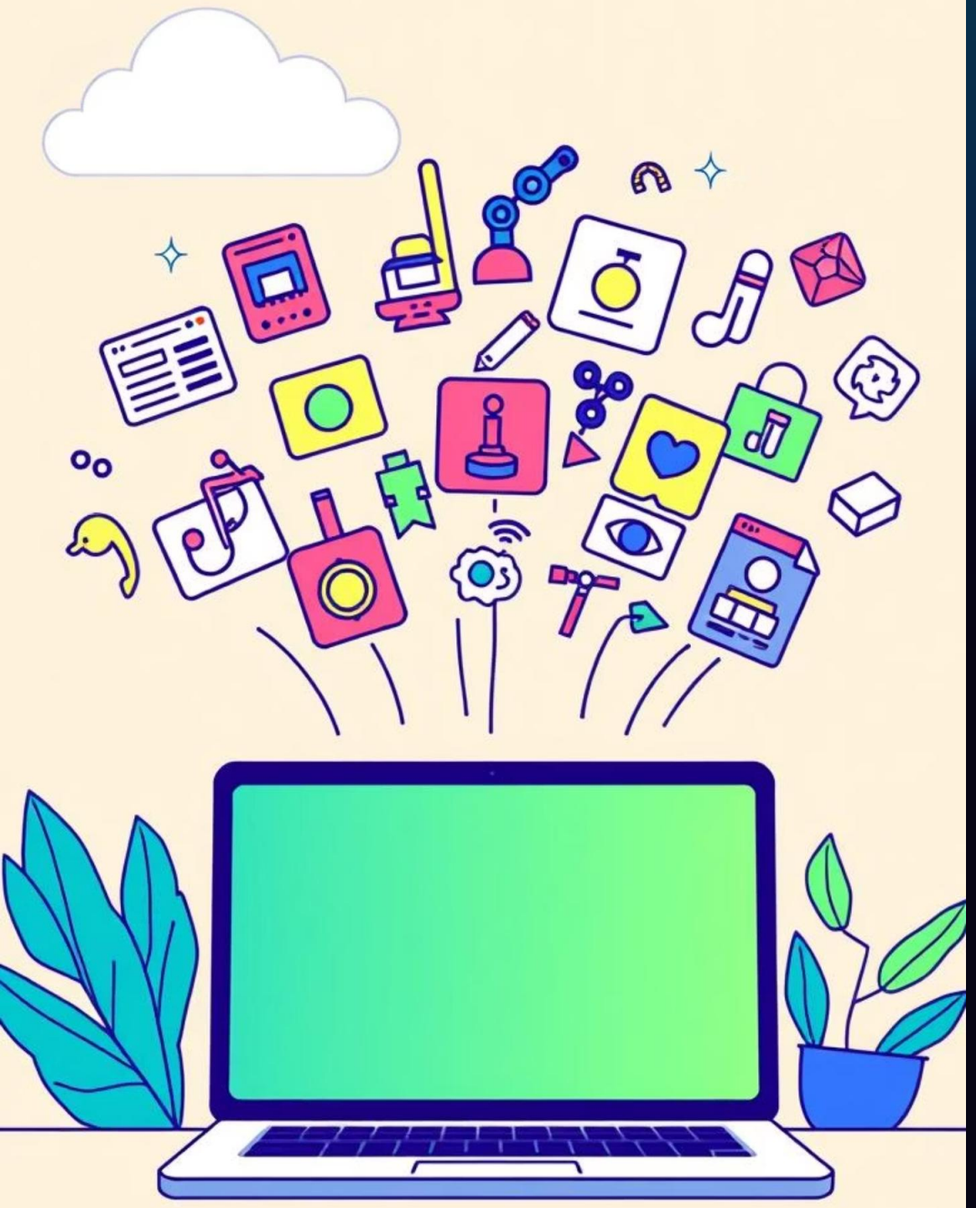
Preparing and understanding the Udemy course dataset.

Modeling

Developing the content-based recommendation engine.

Deployment & Interface

Bringing the system to users with an intuitive Streamlit app.



Dataset Overview: Udemy Courses

Our foundation is a rich dataset of over 3,600 Udemy courses, providing a comprehensive view of online learning content.

- **Source:** Publicly available Udemy courses dataset.
- **Key Features:** Course title, description, subject, level, price, number of subscribers, reviews, and duration.
- **Diversity:** Spanning various subjects from web development to music, offering a broad spectrum for recommendations.

Dataset Overview

A	B	C	D	E	F	G	H	I	J	K	
course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
1070968	Ultimate Investm	https://www.uc	TRUE	200	2147	23	51	All Levels	1.5	2017-01-18T20:58:58Z	Business Finance
1113822	Complete GST Co	https://www.uc	TRUE	75	2792	923	274	All Levels	39	2017-03-09T16:34:20Z	Business Finance
1006314	Financial Modeli	https://www.uc	TRUE	45	2174	74	51	Intermediate Level	2.5	2016-12-19T19:26:30Z	Business Finance
1210588	Beginner to Pro -	https://www.uc	TRUE	95	2451	11	36	All Levels	3	2017-05-30T20:07:24Z	Business Finance
1011058	How To Maximize	https://www.uc	TRUE	200	1276	45	26	Intermediate Level	2	2016-12-13T14:57:18Z	Business Finance
192870	Trading Penny St	https://www.uc	TRUE	150	9221	138	25	All Levels	3	2014-05-02T15:13:30Z	Business Finance
739964	Investing And Tra	https://www.uc	TRUE	65	1540	178	26	Beginner Level	1	2016-02-21T18:23:12Z	Business Finance
403100	Trading Stock Ch	https://www.uc	TRUE	95	2917	148	23	All Levels	2.5	2015-01-30T22:13:03Z	Business Finance
476268	Options Trading	https://www.uc	TRUE	195	5172	34	38	Expert Level	2.5	2015-05-28T00:14:03Z	Business Finance
1167710	The Only Investm	https://www.uc	TRUE	200	827	14	15	All Levels	1	2017-04-18T18:13:32Z	Business Finance
592338	Forex Trading Sec	https://www.uc	TRUE	200	4284	93	76	All Levels	5	2015-09-11T16:47:02Z	Business Finance
975046	Trading Options '	https://www.uc	TRUE	200	1380	42	17	All Levels	1	2016-10-18T22:52:31Z	Business Finance
742602	Financial Manag	https://www.uc	TRUE	30	3607	21	19	All Levels	1.5	2016-02-03T18:04:01Z	Business Finance
794151	Forex Trading Co	https://www.uc	TRUE	195	4061	52	16	All Levels	2	2016-03-16T15:40:19Z	Business Finance
1196544	Python Algo Trad	https://www.uc	TRUE	200	294	19	42	All Levels	7	2017-04-28T16:41:44Z	Business Finance
504036	Short Selling: Lea	https://www.uc	TRUE	75	2276	106	19	Intermediate Level	1.5	2015-06-22T21:18:35Z	Business Finance
719698	Basic Technical A	https://www.uc	TRUE	20	4919	79	16	Beginner Level	1.5	2016-01-08T17:21:26Z	Business Finance
564966	The Complete Ch	https://www.uc	TRUE	200	2666	115	52	All Levels	4	2015-08-10T21:07:35Z	Business Finance
606928	7 Deadly Mistake	https://www.uc	TRUE	50	5354	24	23	All Levels	1.5	2015-09-21T18:10:34Z	Business Finance
58977	Financial Statem	https://www.uc	TRUE	95	8095	249	12	Beginner Level	0.583333333	2013-06-09T00:21:26Z	Business Finance
1242604	Winning Forex Tr	https://www.uc	TRUE	200	809	3	25	All Levels	2	2017-06-06T02:54:04Z	Business Finance
798740	Forex Traders - C	https://www.uc	TRUE	200	2295	84	39	All Levels	4	2016-05-02T19:26:48Z	Business Finance
506568	Create A Busines	https://www.uc	TRUE	75	10149	83	16	All Levels	2	2015-05-26T17:25:46Z	Business Finance

Exploratory Data Analysis (EDA)

understand key characteristics of each course

Column Name	Description
`course_id`	Unique identifier for each course
`course_title`	Title of the course
`url`	Link to the course on Udemy
`is_paid`	Indicates if the course is free or paid (`TRUE` = Paid)
`price`	Course price in USD
`num_subscribers`	Total number of students enrolled in the course
`num_reviews`	Number of student reviews submitted
`num_lectures`	Total number of video lectures in the course
`level`	Course difficulty level (e.g., Beginner, Intermediate, Expert, All Levels)
`content_duration`	Duration of the course content in hours
`published_timestamp`	Date and time when the course was published
`subject`	Category or domain of the course (e.g., Business Finance, Web Development)

Data Cleaning

Removed Duplicates

Ensured each course is unique by checking `course_id` and `course_title`

Handled Missing Values

Checked for nulls in key columns (`price`, `num_subscribers`, `level`, etc.)

Check Outliers

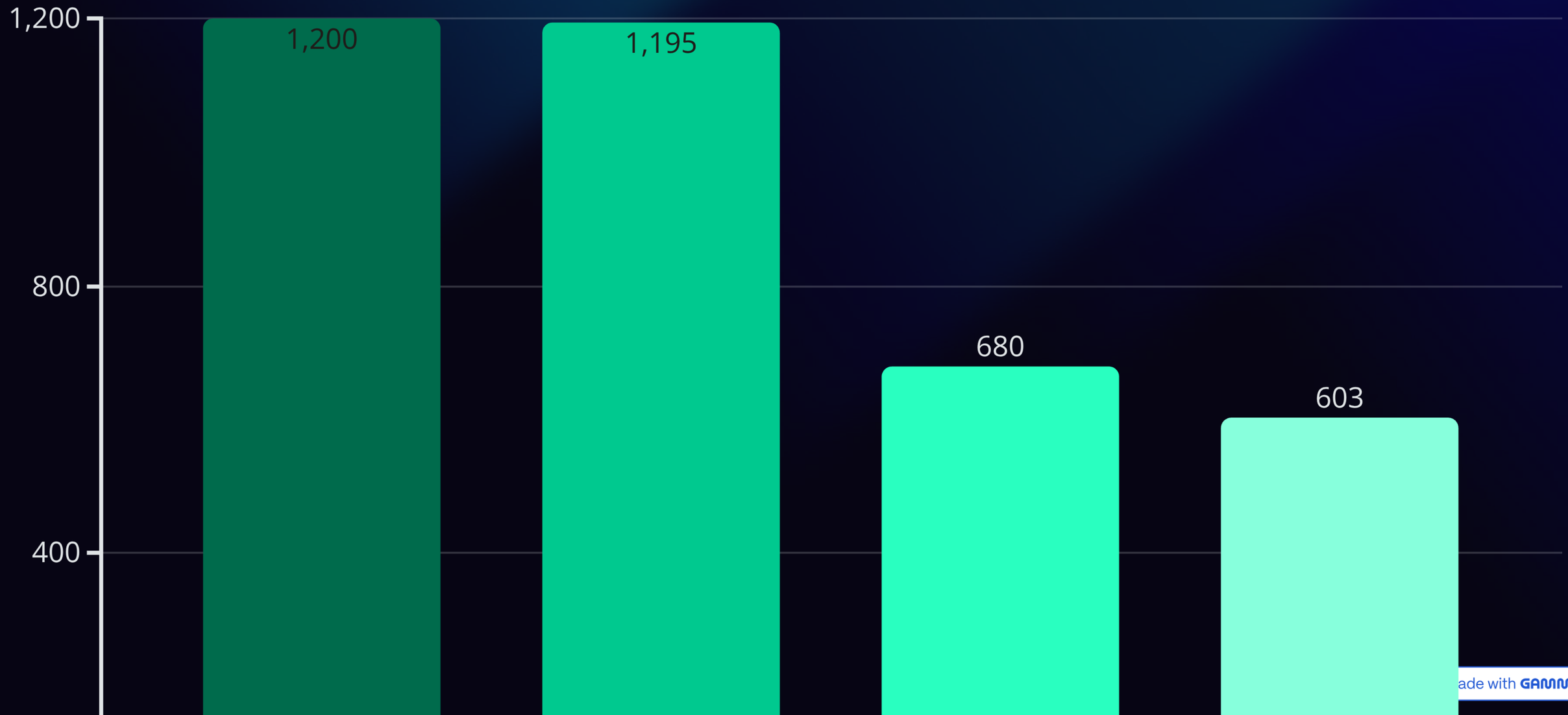
Used IQR to detect/remove unrealistic values

The dataset became **clean**, **consistent**, and **ready for preprocessing and modeling**

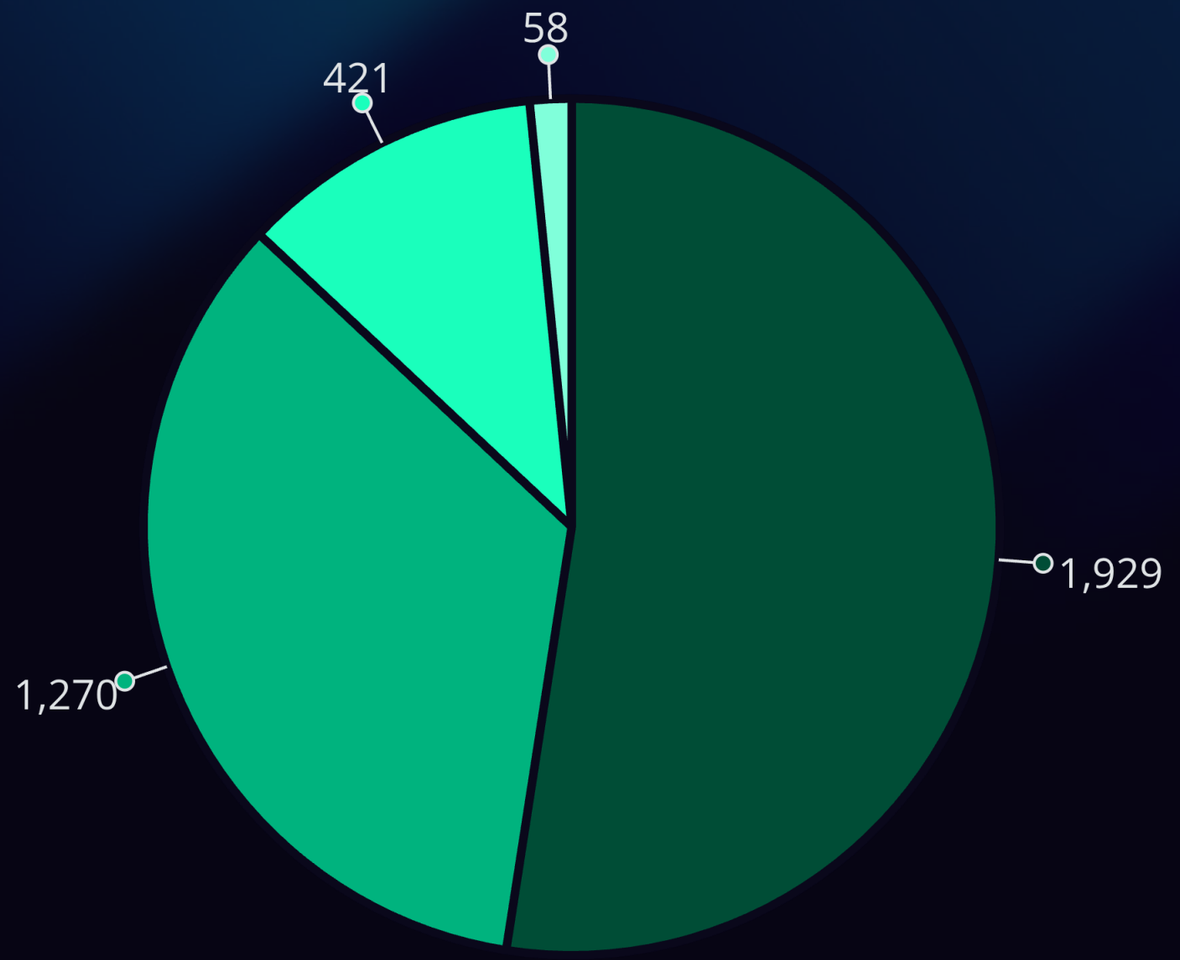
Key Data Visualizations

Visualizing the dataset helps us understand distributions and potential correlations, guiding our feature selection.

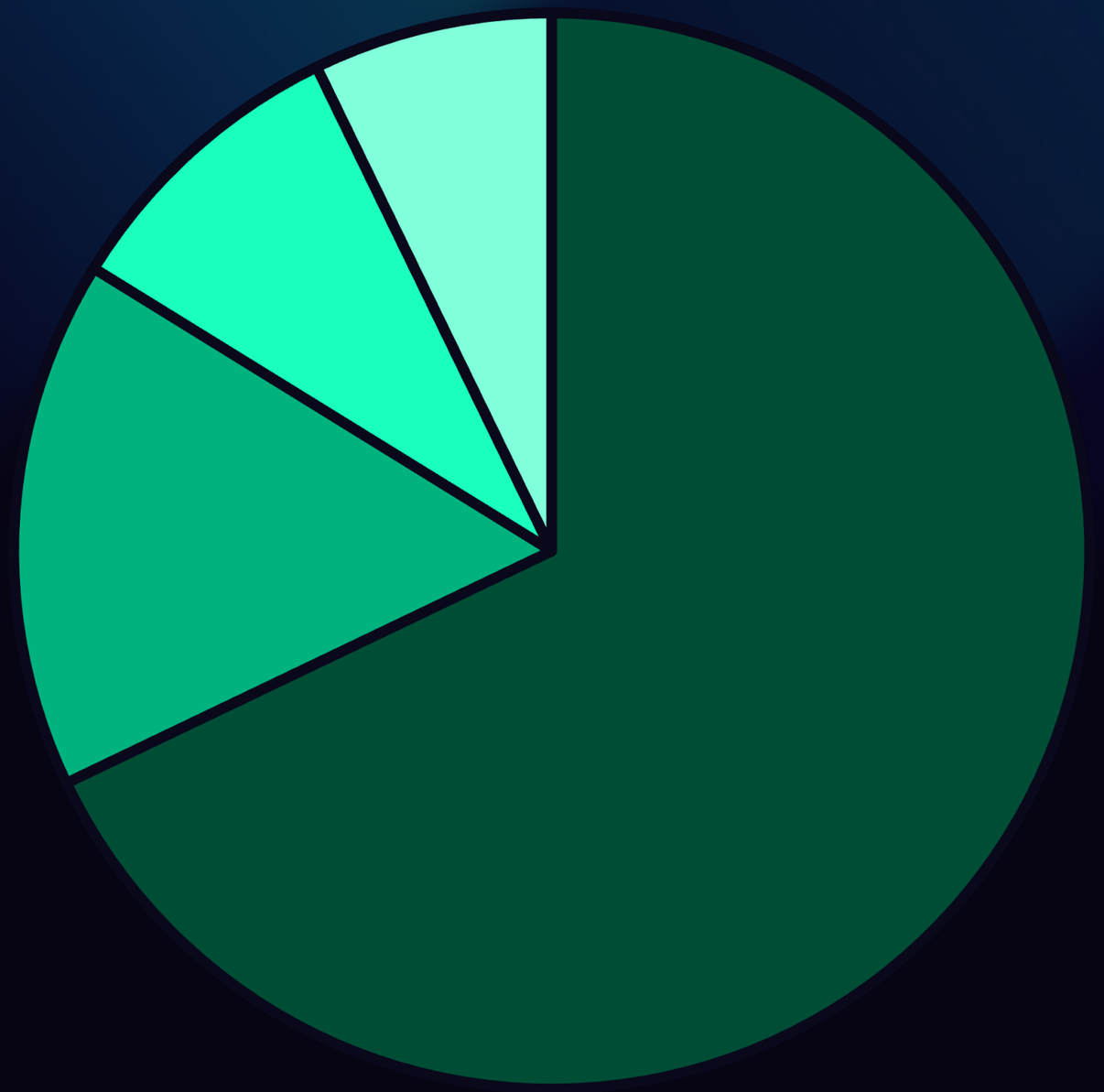
Number of Courses per Subject

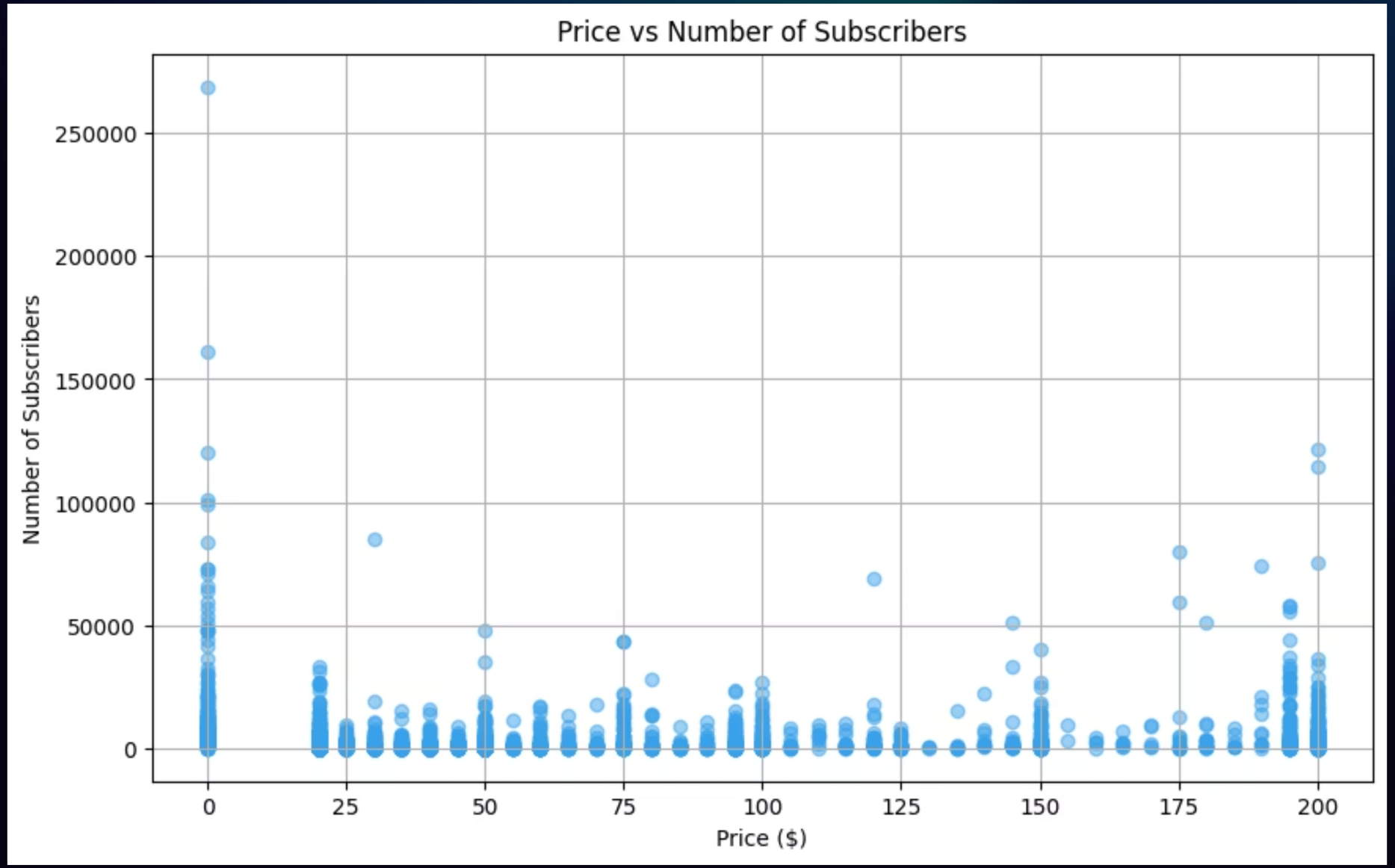


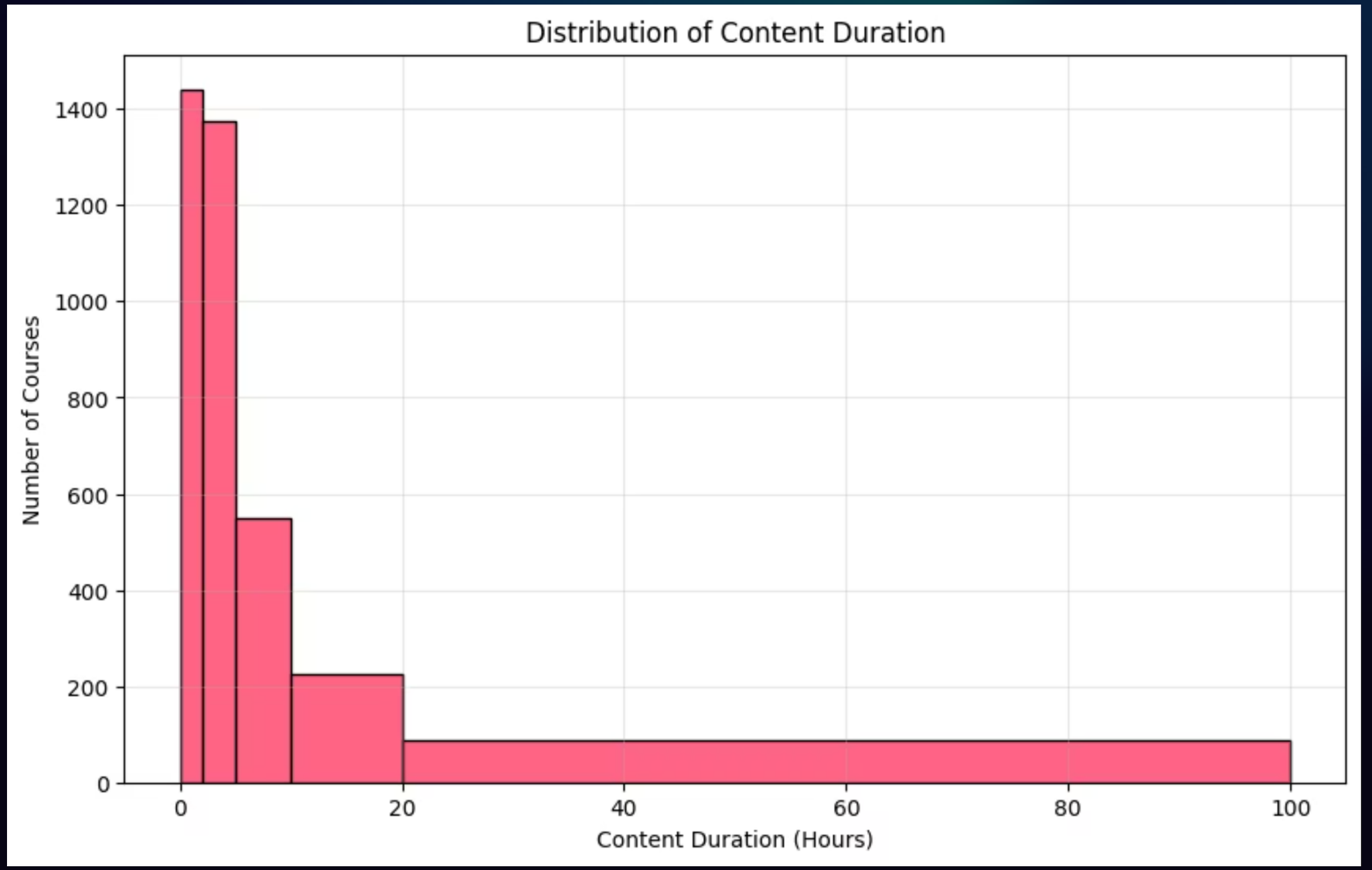
Distribution of Courses by Level

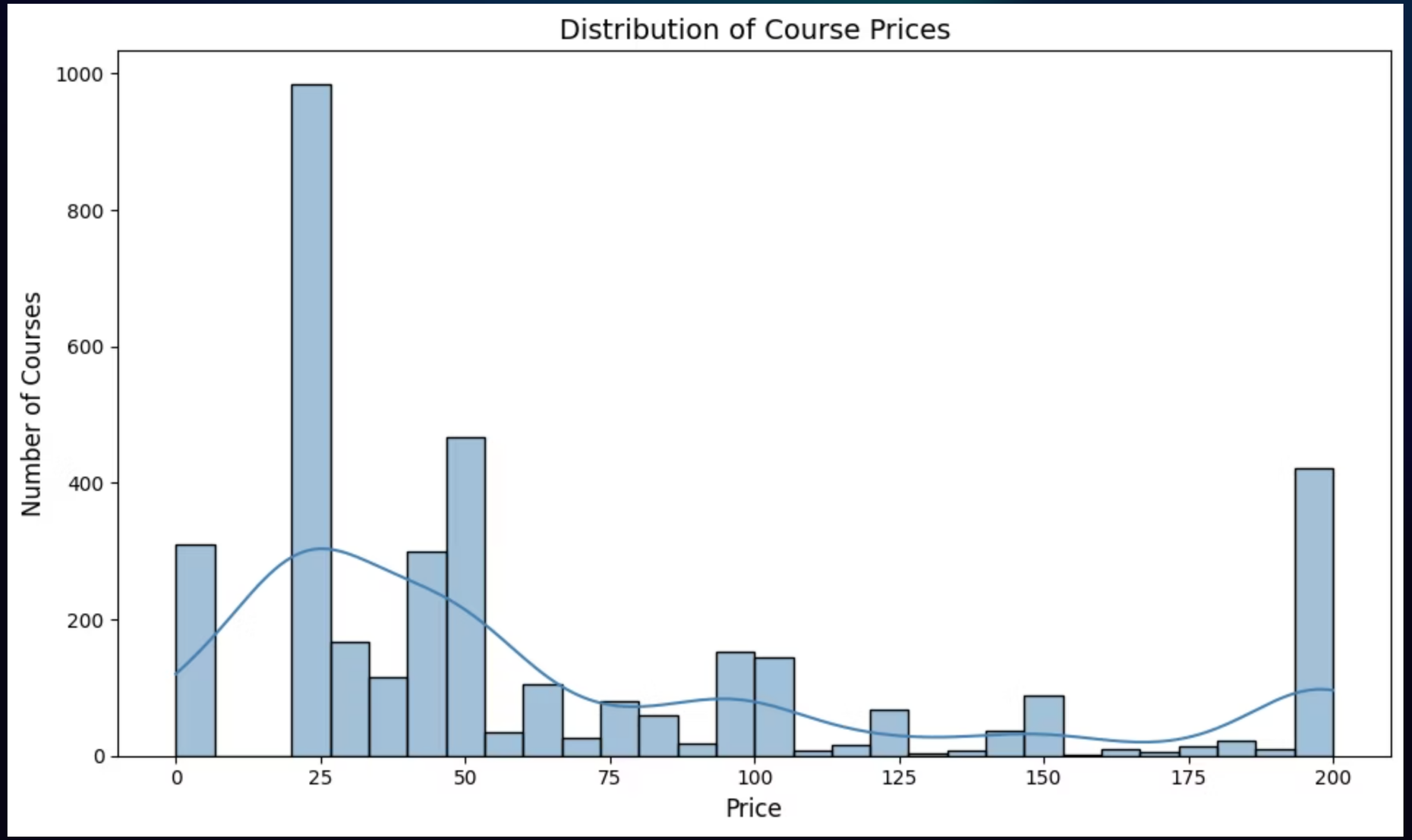


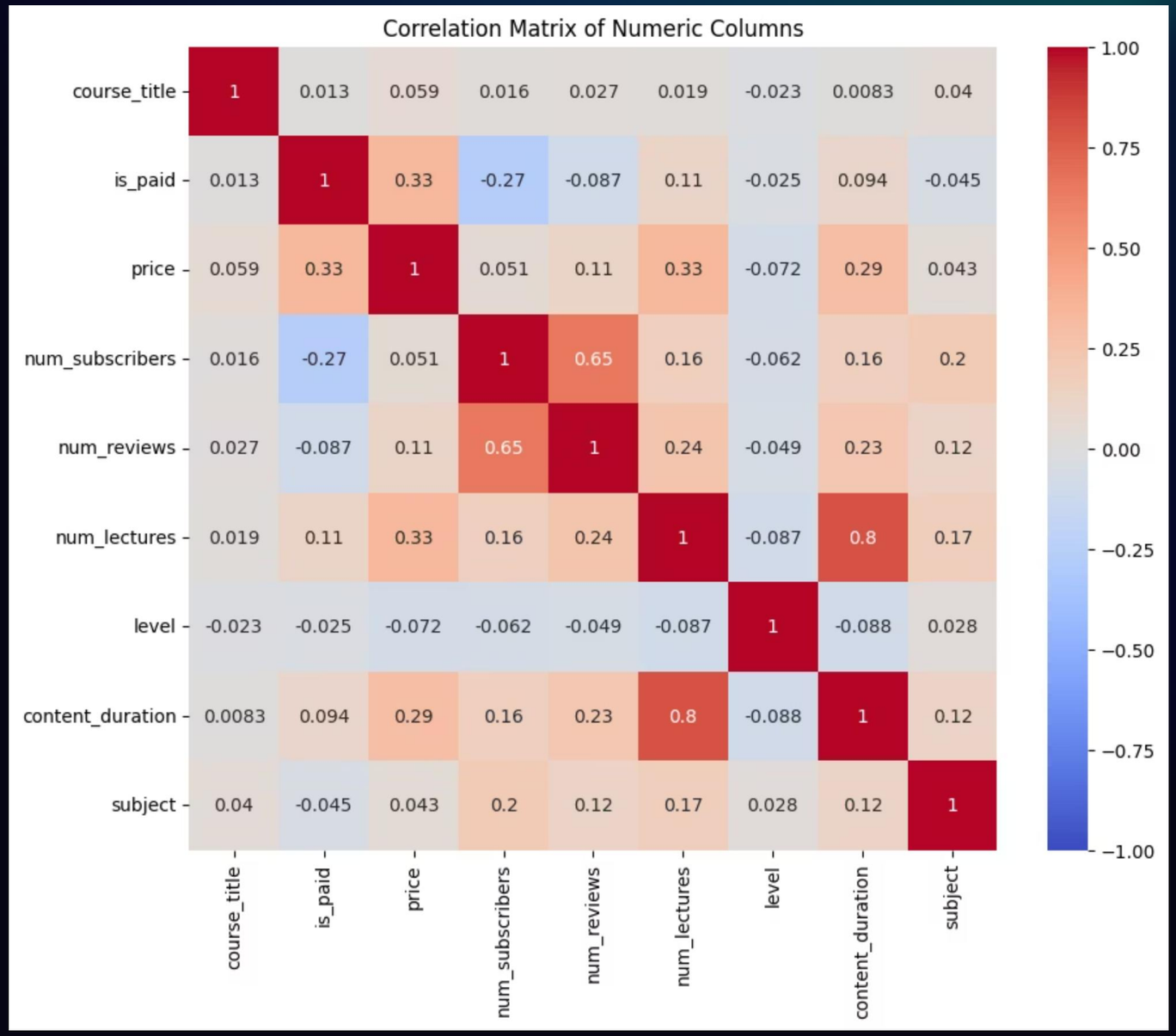
Distribution of Subscribers by Subject











Features Used for Clustering

textual Feature:

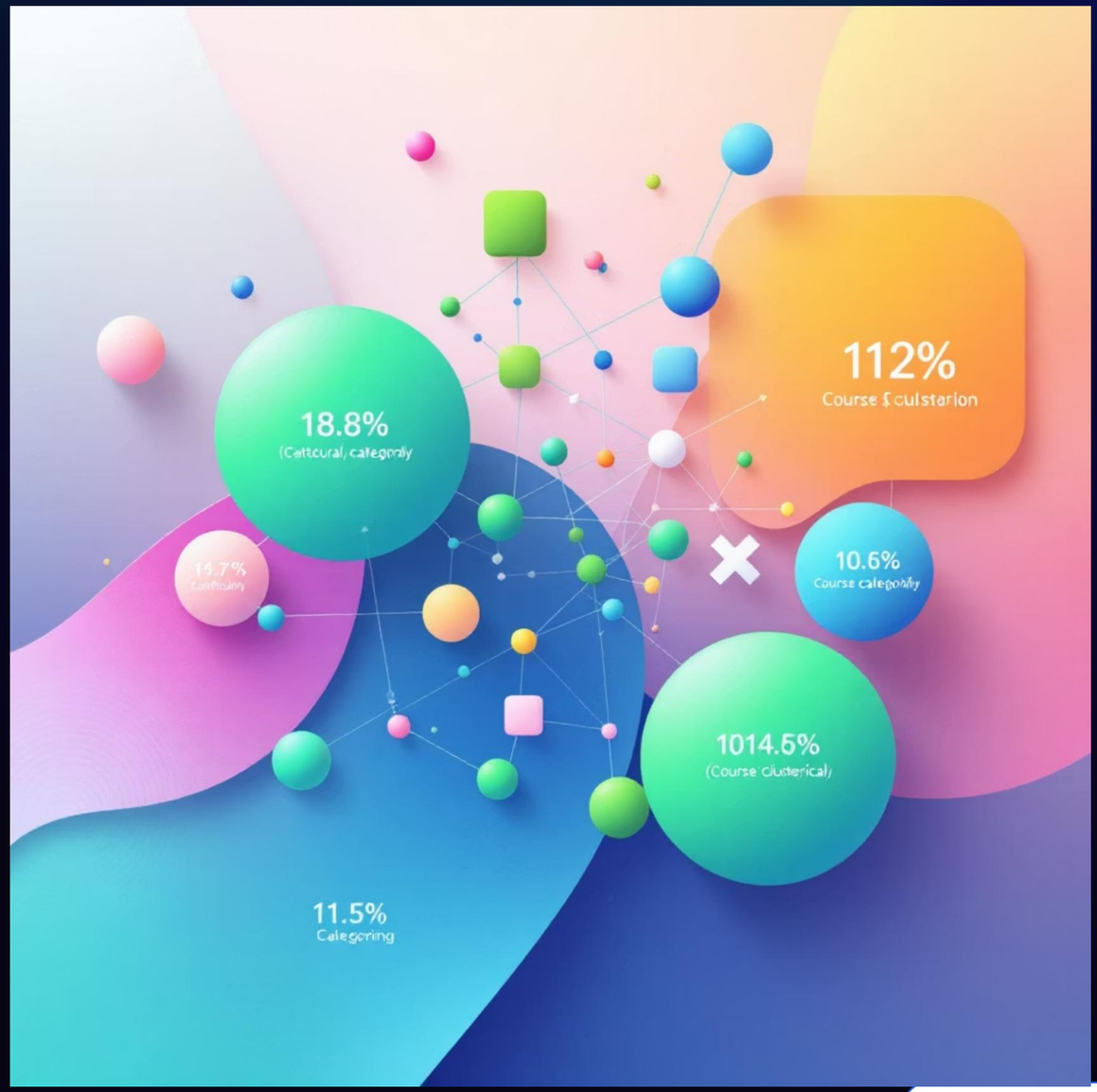
- `course_title` → Processed using **TF-IDF Vectorization**
- Helps capture topic similarities between courses

Categorical Features (Encoded):

- `subject`
- `level`

Numeric Features (Scaled):

- `price`
- `num_subscribers`
- `num_reviews`



Data Preprocessing Pipeline

Transforming raw course data into a structured format for effective modeling is crucial.

Text Handling (TF-IDF)

Course titles and descriptions converted into numerical vectors for semantic understanding.

Categorical Encoding

Subjects and levels are one-hot encoded to represent distinct categories numerically.

Numerical Scaling

Subscriber counts and prices normalized to prevent dominance of larger values.

Modeling: Content-Based Filtering

Our recommender system uses content-based filtering, matching courses based on their inherent characteristics.

1

Core Algorithm

Cosine similarity to measure similarity between course feature vectors.

2

Feature Weighting

Experimented with weighting features like course description (TF-IDF) higher for semantic relevance.

3

Feature Experimentation

Tested combinations of text features (title, description) and numeric features (price, subscribers) for optimal performance.

First, I tried **K-Means** with **PCA** to reduce the dimensionality of the features—like price, number of subscribers, reviews, lectures, and content duration, plus encoded subject and level. I got some pretty good **Silhouette Scores**, which showed the clusters were tight and well-separated. Then, I went a step further and used **embeddings** for the course titles, generated with a model like all-MiniLM-L6-v2, to capture the semantic meaning of the titles. I combined these embeddings with the other features and ran K-Means again.

For the feature weights, I tested two approaches. One, I used **fixed weights**, giving more importance to stuff like num_subscribers and the title embeddings. Two, I ran **Grid Search** to find the best weights for the features, testing different combinations. The Silhouette Scores came out really strong—honestly, they looked great, sometimes above 0.6, which is solid for clustering.

But, here's the thing: when I tested the clusters in a practical setting—like recommending courses based on them—the results weren't that accurate. Some predictions were off, and the recommendations didn't always make sense, like suggesting courses that weren't really similar.

So, I decided to try **DBSCAN** instead, thinking it might handle outliers better or find more natural clusters. I used the same setup: PCA in one run, embeddings in another, fixed weights, and Grid Search for weights, plus tuning eps and min_samples. Again, the Silhouette Scores were good, especially with the best parameters from Grid Search. But, same problem—when I applied it practically, the recommendations were still off. Some courses ended up in the noise cluster, and others didn't group as expected.

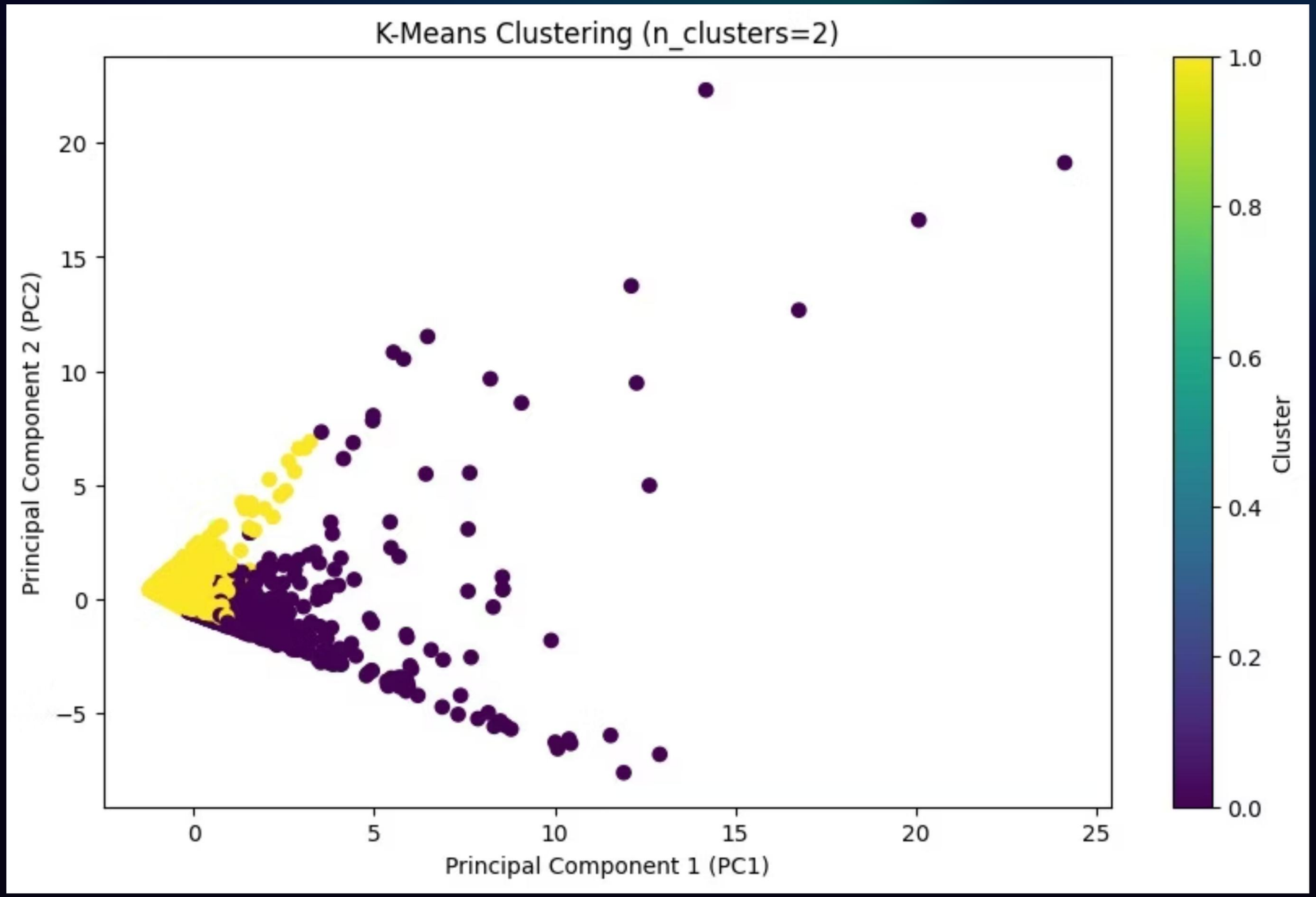
Here's a sample of the results: for **K-Means** with **PCA** , I got a Silhouette Score around 0.5449 with 2 clusters. For **DBSCAN**, the run had a Silhouette Score of about 0.567 with 3 clusters . But in both cases, when I checked the actual course recommendations, they didn't align well with what users would expect.

Appling Kmean With PCA

```
Explained Variance Ratio: [0.45419456 0.27698876 0.16038283 0.06905668 0.03937718]

Grid Search Results:
   n_clusters  silhouette_score       inertia
0           2          0.544908  14192.473969
1           3          0.506211  10787.611718
2           4          0.508255   7876.805244
3           5          0.522939   6751.834561
4           6          0.339572   6252.126414
5           7          0.366010   5004.939774
6           8          0.375513   4681.739173
7           9          0.380105   4492.191780
8          10          0.413572   4171.267535

Best Parameters:
n_clusters: 2
Best Silhouette Score: 0.5449084797102958
```

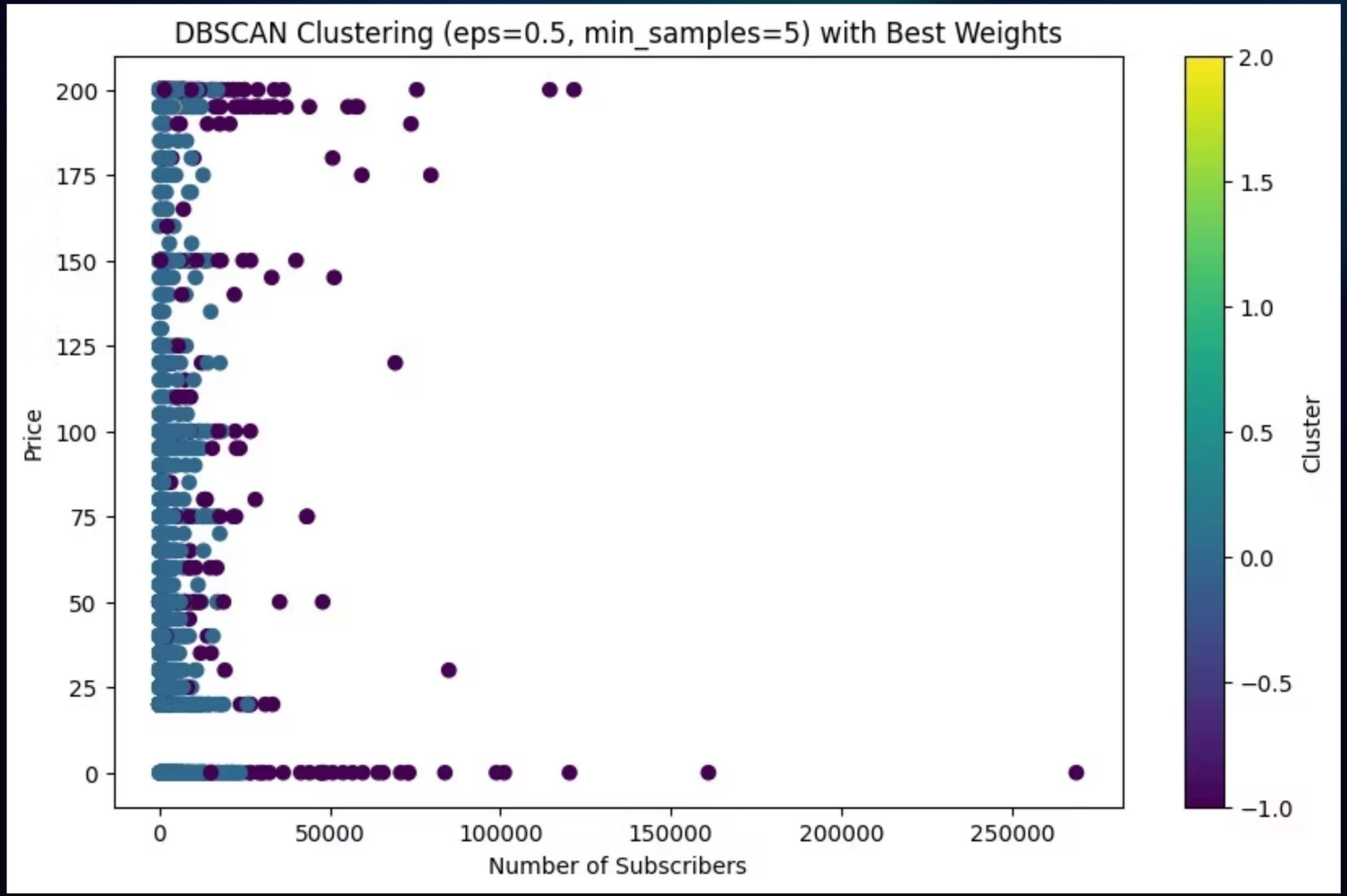


Appling DBSCAN With grid search for Weights and parameters

```
Grid Search Results (Top 5 by Silhouette Score):
    eps  min_samples                                                 weights  \
4   0.5            5  {'price': 0.5, 'num_subscribers': 1.0, 'num_re...
8   0.7            7  {'price': 0.5, 'num_subscribers': 1.0, 'num_re...
7   0.7            5  {'price': 0.5, 'num_subscribers': 1.0, 'num_re...
17  0.7            7  {'price': 0.5, 'num_subscribers': 1.0, 'num_re...
16  0.7            5  {'price': 0.5, 'num_subscribers': 1.0, 'num_re...

    n_clusters  n_noise  silhouette_score
4            3      274          0.567055
8            2      198          0.562350
7            2      182          0.560877
17           2      229          0.560262
16           2      203          0.556389

Best Parameters:
eps: 0.5
min_samples: 5
Weights: {'price': 0.5, 'num_subscribers': 1.0, 'num_reviews': 1.0, 'num_lectures': 1.0, 'content_duration': 1.0}
Best Silhouette Score: 0.5670546299913514
Number of Clusters: 3
Number of Noise Points: 274
```



But here’s the problem: when I tested these models practically—like trying to recommend courses based on the clusters—the predictions were off. For example, with K-Means, I’d get recommendations that didn’t really match the course’s topic or user expectations. Same thing with DBSCAN; some courses ended up in the noise cluster, and others were grouped in ways that didn’t make sense practically.

So, I switched to **Cosine Similarity** for recommendations. I used the same features—price, subscribers, reviews, and the title embeddings—and calculated the similarity between courses. This approach worked much better! The recommendations were more accurate, like suggesting Python courses for a Python course, or matching based on price and popularity. For instance, when I tested Cosine Similarity on a course, it recommended courses with similar titles and features, and they felt way more relevant compared to K-Means or DBSCAN.

Here’s a sample of the results:

Feature Weighting

To control the influence of each feature (textual, numeric, categorical) on the recommendation results — making the similarity more balanced and meaningful.

🧩 Feature Groups & Weights

Feature Group	Features Included	Applied Weight
Textual	`course_title` (TF-IDF vectorized)	🟢 High
Categorical	`subject`, `level` (one-hot encoded)	🟡 Medium
Numeric	`price`, `num_subscribers`, `num_reviews` (scaled)	🔵 Low

Choosed Course

	123
course_id	1100054
course_title	FOREX TRADING - Learn in a quick + profitable ...
url	https://www.udemy.com/forex-the-only-simple-tr...
is_paid	True
price	100
num_subscribers	271
num_reviews	48
num_lectures	47
level	All Levels
content_duration	3.5
blished_timestamp	2017-02-20T23:45:20Z
subject	Business Finance

Choosed Course

Cosine Similarity (Original):

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	similarity
7	964828	Forex Trading - Learn An Effective Forex Tradi...	https://www.udemy.com/forex-trading-strategy-t...	True	95	223	34	16	All Levels	1.5	2016-12-14T22:45:02Z	Business Finance	0.955458
9	324656	Forex Trading	https://www.udemy.com/how-to-trade-forex/	True	95	136	14	25	All Levels	4.5	2014-11-18T10:52:45Z	Business Finance	0.953922
3	575476	Forex trading made simple	https://www.udemy.com/forex-mentor-online-school/	True	100	45	4	31	All Levels	1.5	2015-09-25T19:08:09Z	Business Finance	0.947554

Cosine Similarity

KMeans Recommendation (PCA):

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
34	1041926	How to Trade Forex like a Hedge Fund: Long FX ...	https://www.udemy.com/how-to-trade-forex-like-...	True	115	504	5	32	All Levels	1.5	2017-03-23T16:23:33Z	Business Finance
4	876646	Direction-Independ Trading - Elite Forex Trade...	https://www.udemy.com/direction-independ-tradi...	True	120	973	15	28	All Levels	2.5	2016-06-21T22:46:41Z	Business Finance
15	1189592	Learn How To Successfully Trade Forex: In 5 Si...	https://www.udemy.com/learn-how-to-successfull...	True	95	64	7	44	All Levels	8.5	2017-05-03T22:09:03Z	Business Finance

Kmeans with PCA

	course_id	course_title	url	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
10	944534	Forex:Trade Management & Psychology	https://www.udemy.com/forexmoney-management-ps...	True	150	13	0	19	Intermediate Level	2.0	2016-09-21T19:07:06Z	Business Finance
6	1247394	Cryptocurrency Trading: Complete Guide To Trad...	https://www.udemy.com/cryptocurrency-trading/	True	95	367	42	35	All Levels	5.0	2017-06-21T23:18:47Z	Business Finance
72	581598	Commodity Futures Day Trading Strategies	https://www.udemy.com/futures-day-trade-course/	True	95	91	5	28	All Levels	5.0	2015-08-28T23:47:29Z	Business Finance

DBSCAN With PCA

Deployment with Streamlit

The recommender system is deployed as an interactive web application using Streamlit, providing a seamless user experience.

Ready to find your next course?

Let’s Try the Interface

Key Takeaways & Next Steps

This project demonstrates a robust approach to personalized course recommendations.

Personalized Learning

Content-based filtering provides highly relevant course suggestions.

Scalable Deployment

Streamlit offers a rapid and interactive deployment solution.

Future Enhancements

Incorporating collaborative filtering and real-time user feedback for even smarter recommendations.