

DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING
PROJECT REPORT

(Project Semester January-April 2025)

ANALYSIS ON UDEMY COURSES

Submitted by

Bhoomi

Registration No. 12308951

Programme and Section B.Tech. (Computer Science) , KM007
Course Code INT375

Under the Guidance of

Dr. Mrinalini Rana(UID:-22138)

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Bhoomi bearing Registration no. 123089851 has completed INT375 project titled, “**Analysis on udemy courses**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science

Lovely Professional University

Phagwara, Punjab.

Date: 09-04-2025

DECLARATION

I, Bhoomi, student of B.tech. under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 09-04-2025

Signature

Registration No. 12308951

Bhoomi

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who supported me during the development of this agricultural dashboard project.

A special thanks to my mentor/faculty Dr. Mrinalini Rana for their invaluable guidance and motivation. I also appreciate my institution for providing the necessary resources and a great environment for this project.

I am grateful to my team and peers for their constructive suggestions and assistance with technical challenges, as well as to the tools like Microsoft Excel and Power Query that helped transform raw data into an interactive dashboard.

Finally, I acknowledge the open data sources that provided relevant information for analysis. This project has been a valuable learning experience, and I am truly thankful for everyone's contributions.

TABLE OF CONTENTS

INTRODUCTION.....	6
SOURCE OF DATASET	7
DATA PREPROCESSING	7
ANALYSIS ON DATASET	9
OBJECTIVE 1: FREE VS PAID COURSES BY SUBJECT	9
OBJECTIVE 2: OPTIMAL COURSE DURATION	10
OBJECTIVE 3: MOST POPULAR COURSE TITLES.....	12
OBJECTIVE 4: COURSE REVIEWS BY DIFFICULTY LEVEL	13
OBJECTIVE 5: TRENDS IN COURSE CREATION OVER TIME	15
OBJECTIVE 6: COURSES PUBLISHED BY DAY OF THE WEEK.....	16
CONCLUSION	17
FUTURE SCOPE.....	18
REFERENCES.....	19
SCREENSHOTS:.....	20
LINKEDIN SCREENSHOTS:	20
GITHUB SCREENSHOTS:.....	21

INTRODUCTION

In the modern era of digital transformation, data has become a crucial asset in driving decisions across industries, especially in the field of education and online learning. With the rapid rise of e-learning platforms like Udemy, Coursera, and edX, a massive amount of data is generated every day in the form of course enrollments, user engagement, reviews, pricing trends, and more. Among these platforms, Udemy stands out as one of the largest marketplaces for online courses, offering educational content across a broad spectrum of subjects to learners around the world.

This project was carried out as part of a Skill-Based Assignment aimed at developing and evaluating technical proficiency in Python for data manipulation, analysis, and visualization. The primary goal was to gain hands-on experience with a real-world dataset and apply essential data science techniques to extract meaningful insights. The dataset used for this project is titled "Udemy Courses", which contains detailed information about various courses hosted on the Udemy platform. It includes attributes such as course title, subject, level, price, number of subscribers, reviews, content duration, publication date, and more.

The analysis began with a thorough data cleaning process to handle missing values, correct inconsistencies, and ensure that the dataset was ready for meaningful analysis. This was followed by Exploratory Data Analysis (EDA), where both univariate (single variable) and bivariate (relationship between two variables) analyses were performed to uncover patterns, distributions, and correlations within the data. Python libraries like pandas, numpy, matplotlib, and seaborn were extensively used to structure the data and create insightful visualizations.

To make the analysis more engaging and innovative, I formulated six unique and creative problem statements that reflected real-world scenarios and decision-making needs, such as understanding the most engaging course categories, exploring pricing strategies for free vs paid courses, examining trends over the years (especially the impact of events like COVID-19), and identifying factors that influence the popularity of a course. For each problem, appropriate visualizations were designed—including bar charts, pie charts, histograms, scatter plots, and line graphs—to communicate the findings clearly and effectively.

What sets this project apart is not just the technical implementation but also the effort to tie every analysis back to a real-world objective, making the insights actionable. The use of storytelling through data, backed by strong visuals, helped in turning raw information into knowledge.

Overall, this project served as an excellent opportunity to strengthen core data science skills, enhance understanding of Python-based data workflows, and gain practical exposure to solving problems through data. It reflects my personal journey in approaching a data science task—from dataset selection to insight presentation—and highlights the immense potential of data in shaping strategies in the online education industry.

SOURCE OF DATASET

[Datasets/winequality-white.csv at master · MainakRepositor/Datasets](#)

DATA PREPROCESSING

To extract meaningful insights from the Udemmy Courses dataset, it was essential to first prepare and understand the data. This involved a systematic process that included data cleaning, preprocessing, and exploratory data analysis (EDA). Below are the detailed steps followed in this project:

Step 1: Initial Loading and Inspection of the Dataset

The dataset was first imported and examined to understand its structure and overall quality. This involved viewing the number of records and features, checking data types, and scanning the top few rows for a quick glance at the contents. Summary statistics such as mean, median, and standard deviation were also reviewed to identify any glaring inconsistencies or abnormalities.

Step 2: Handling Missing Values

A null value check was performed to identify any incomplete records. Although the dataset was mostly clean, a few rows had missing values in less relevant columns. These records were dropped to maintain consistency and ensure the accuracy of further analysis. This step helped prevent distortions in the analysis caused by incomplete data.

Step 3: Removing Duplicate Entries

Duplicate course records were identified and removed. These may occur due to repeated listings or data entry errors and can significantly distort analysis by over-representing certain entries. Removing duplicates ensured that each course was represented only once, resulting in more precise statistical interpretations.

Step 4: Converting Data Types and Extracting Time-Based Features

Several columns required data type conversion. For example, the `published_timestamp` was converted from string format to a datetime object. From this, the year of publication was extracted and stored in a new column. This made it easier to analyze trends over time and helped in grouping the data more effectively for time-based visualizations.

Step 5: Formatting and Simplifying Column Values

Column values were formatted for clarity and consistency. The `is_paid` column was converted from a binary indicator to readable labels like "Free" and "Paid," making the dataset more understandable. Standardizing column names and values also contributed to a smoother and more interpretable visualization experience.

Step 6: Detecting and Handling Outliers using the Z Score

Outlier detection was performed on numerical columns such as `price`, `num_subscribers`, and `content_duration`. Using the Z Score method, values significantly above or below the normal range were identified. Although some values were extreme, they were not removed as they represented valid real-world scenarios, such as highly successful or premium courses. Instead, they were carefully considered during visualizations to avoid distortion.

Step 7: Correlation Analysis Between Numerical Features

A correlation matrix was created to identify how numerical variables such as `price`, `number of reviews`, and `number of subscribers` relate to each other. Visualized using a heatmap, this helped us spot patterns such as strong positive correlations between `subscribers` and `reviews`. These relationships helped in understanding what factors might contribute to a course's popularity.

ANALYSIS ON DATASET

Objective 1: Free vs Paid Courses by Subject

Compare how many free and paid courses are offered in each subject area.

i. General Description

This objective aims to explore how the distribution of free and paid courses varies across different subject areas on Udemy. It provides insight into the availability of free learning resources in various domains, helping users understand which subjects offer more free content and which are largely monetized. This analysis is useful for both learners seeking cost-effective education and instructors deciding pricing strategies for their courses.

ii. Specific Requirements

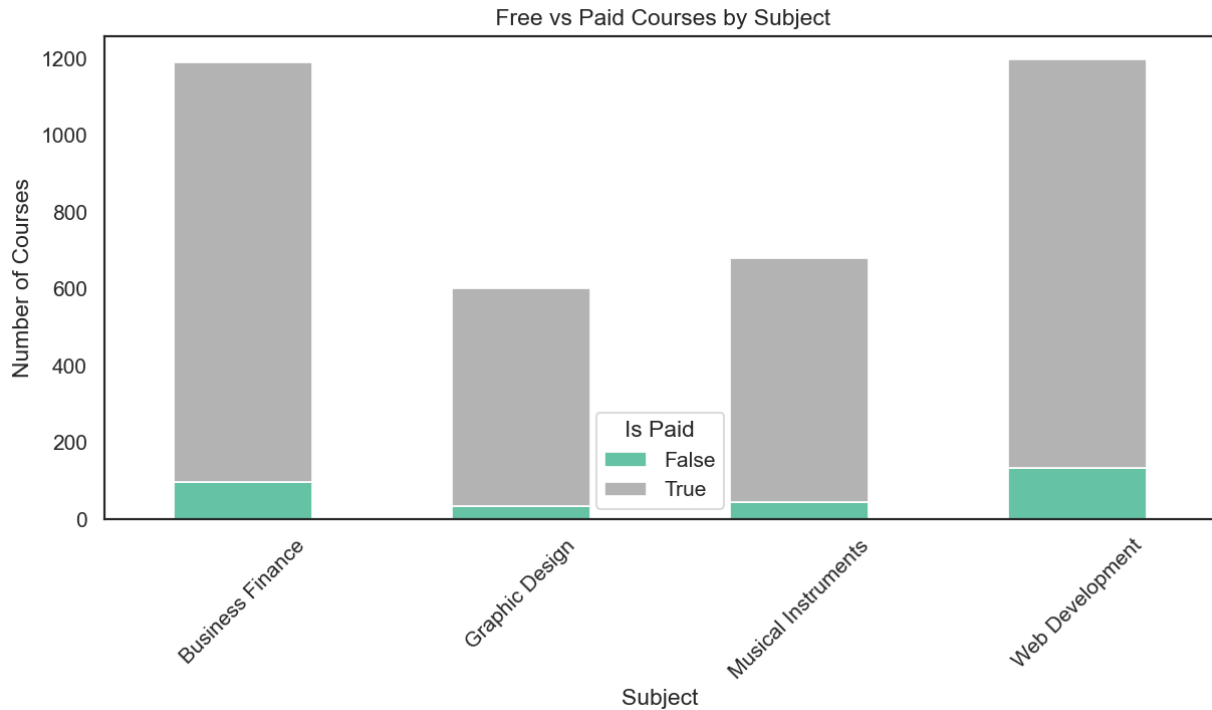
- The `is_paid` column is used to distinguish between free and paid courses.
- The `subject` column is used to categorize the courses by topic.
- The data is grouped by both `subject` and `is_paid` to count how many courses fall into each pricing type within each subject.

iii. Analysis Results

The results show a significant imbalance in the distribution of free and paid courses across subjects. In most subjects, paid courses dominate, especially in Web Development and Business Finance. However, some categories like Musical Instruments and Graphic Design show a relatively higher proportion of free courses. This suggests that while monetization is strong across all categories, some subjects are more open to providing free learning opportunities.

iv. Visualization

A grouped bar plot was used for clear comparison between free and paid courses across each subject area.



This visualization makes it easy to compare the distribution of course types within each subject, providing immediate insight into how pricing models vary across different educational domains.

Objective 2: Optimal Course Duration

Find the ideal course length that attracts the most students.

i. General Description

The goal of this objective is to determine the course duration that tends to attract the highest number of students. In the world of online education, understanding learner preferences regarding course length is crucial. Some users may prefer short, focused content, while others may be more inclined toward in-depth, long-format courses. Identifying this “sweet spot” helps course creators design offerings that are better aligned with student engagement patterns.

ii. Specific Requirements

- The `content_duration` column, which represents the length of a course (in hours), is used to analyze course length.

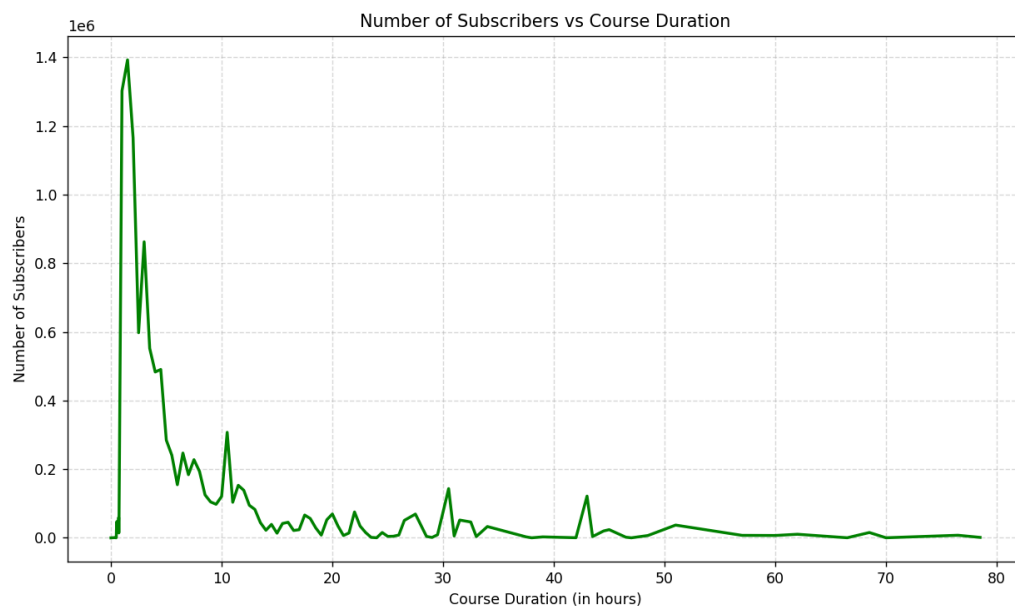
- The num_subscribers column is used to assess popularity and engagement.
- A line plot is created after grouping the data by content duration and summing up the number of subscribers for each unique duration value.
- Duration values are binned (if needed) to avoid clutter and better represent trends.

iii. Analysis Results

The analysis revealed that most students tend to enroll in courses that are **between 1 and 3 hours long**, with another noticeable spike around **10–15 hours**. Extremely short or overly long courses generally received fewer subscribers. This suggests that users may prefer moderately detailed content that's neither too brief nor too time-consuming. Course creators can use this insight to plan their content around these optimal durations to maximize engagement.

iv. Visualization

A **line plot** was used to visualize the trend of course duration versus the number of subscribers.



This visualization clearly illustrates the relationship between course length and student interest. It helps to pinpoint the most effective duration range, which can be a valuable decision-making factor for instructors aiming to boost enrollment.

Objective 3: Most Popular Course Titles

Identify course titles that attracted the highest number of students.

i. General Description

This objective focuses on identifying the specific course titles that have gained the most traction among learners on Udemy. By examining the most popular courses, we can better understand what topics, formats, or even naming styles tend to resonate with the audience. This insight is especially useful for new instructors aiming to structure and name their courses for maximum appeal.

ii. Specific Requirements

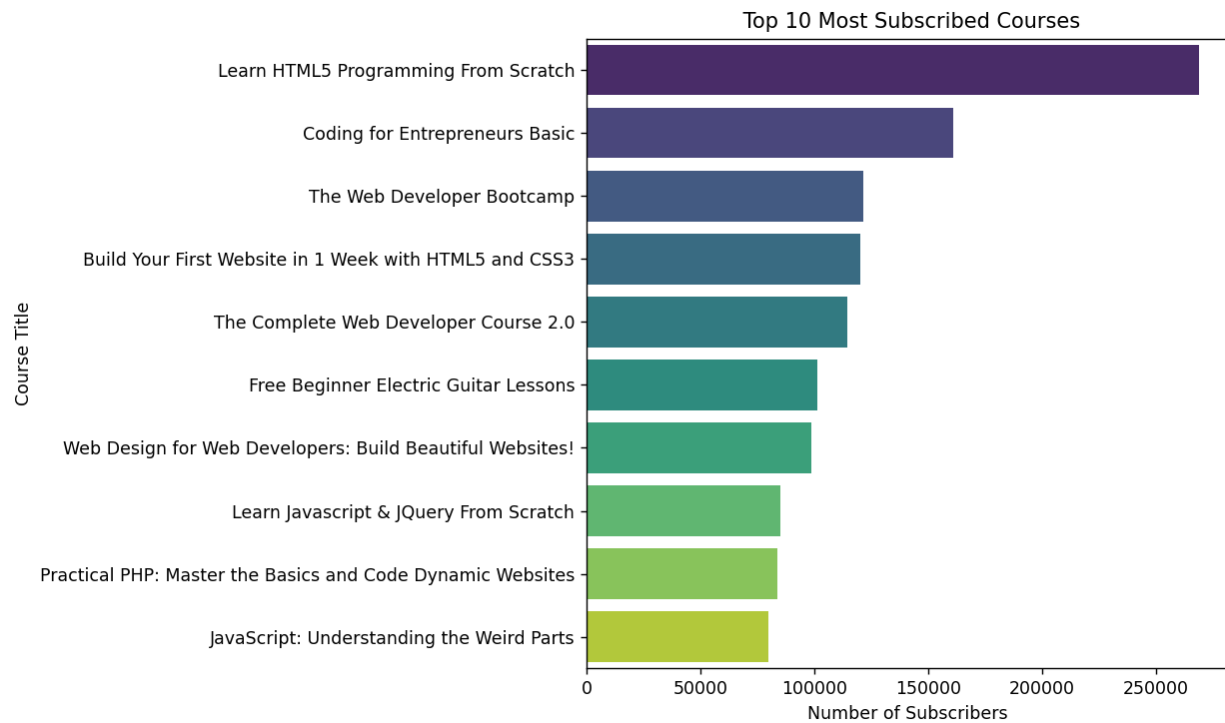
- The `course_title` column is used to list all courses.
- The `num_subscribers` column is used to measure popularity based on total enrollments.
- The data is sorted in descending order of the number of subscribers to identify the top courses.
- A bar plot is used to visually highlight the most popular titles.

iii. Analysis Results

The analysis revealed that the top-performing courses generally focus on technical skills such as **Web Development**, **Python Programming**, and **Data Science**. These courses tend to have more precise, action-oriented titles like “Learn Python Programming Masterclass” or “The Complete Web Developer Course 2.0.” It was also noticed that highly subscribed courses often contain keywords like “Complete,” “Masterclass,” or specific job roles like “Developer” and “Analyst.” These patterns suggest that learners are attracted to comprehensive and career-relevant content.

iv. Visualization

A **horizontal bar plot** was used to clearly display the top 10 most popular course titles.



This visualization effectively communicates which individual courses dominate in popularity, offering valuable inspiration for course creation and marketing.

Objective 4: Course Reviews by Difficulty Level

Analyse which course level (Beginner, Intermediate, etc.) gets the most reviews.

i. General Description

This objective explores how the number of course reviews varies across different difficulty levels, such as **Beginner**, **Intermediate**, **Expert**, and **All Levels**. Understanding this distribution helps reveal which user skill levels tend to be more engaged or more likely to leave feedback. It can also indicate which audience segment is the most active in reviewing and evaluating content.

ii. Specific Requirements

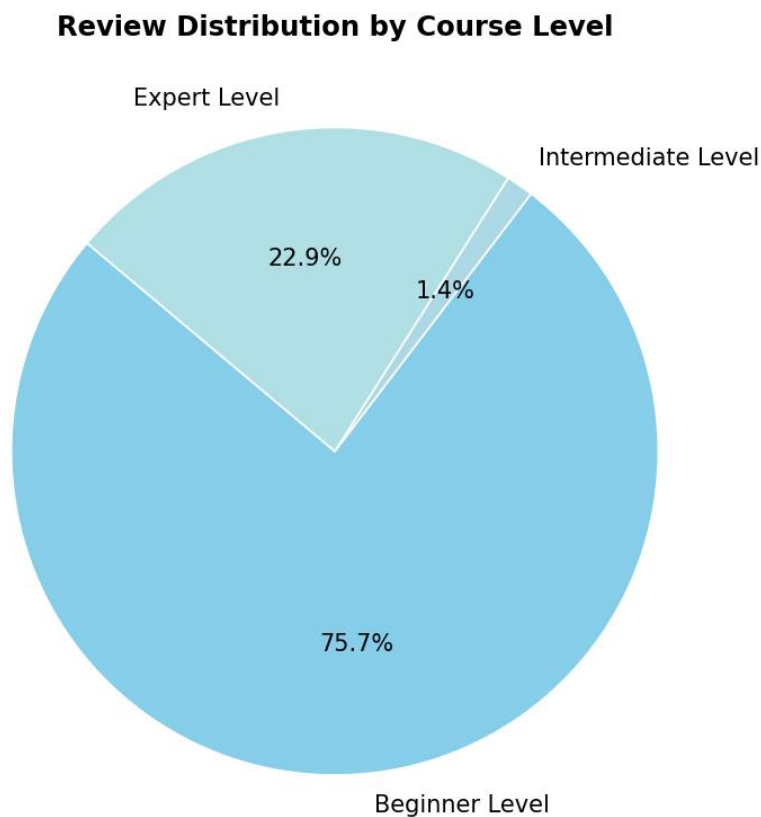
- The level column is used to categorize courses based on their difficulty.
- The num_reviews column is used to measure learner engagement via reviews.
- The data is grouped by level and aggregated using the sum of reviews.
- A clean bar plot is used to visualize total reviews per difficulty level.

iii. Analysis Results

The analysis showed that **Beginner level courses** received the highest number of reviews, followed closely by **All Levels**. This indicates that beginner-friendly content tends to receive more interaction from users, possibly because beginners are more eager to share their learning experience or seek social validation. **Expert level** courses received the least number of reviews, suggesting lower participation or a smaller audience size for highly advanced material.

iv. Visualization

A **pie plot** was used to compare the number of reviews across different difficulty levels clearly and cleanly.



This visualization clearly shows which course levels receive the most feedback, guiding instructors in designing courses tailored to the most responsive audiences.

Objective 5: Trends in Course Creation Over Time

Track how course uploads have changed over the years.

i. General Description

This objective focuses on understanding the pattern of course creation activity on the platform across different years. By examining how the number of courses uploaded has changed over time, we can identify specific years that saw growth or decline in content generation. Such insights help in correlating external events or internal platform strategies with spikes or dips in course creation.

ii. Specific Requirements

- Extract the **year** from the course **published timestamp**.
- Count the number of **courses published per year**.
- Handle any invalid or missing date entries.
- Visualize the data to show clear year-wise trends.

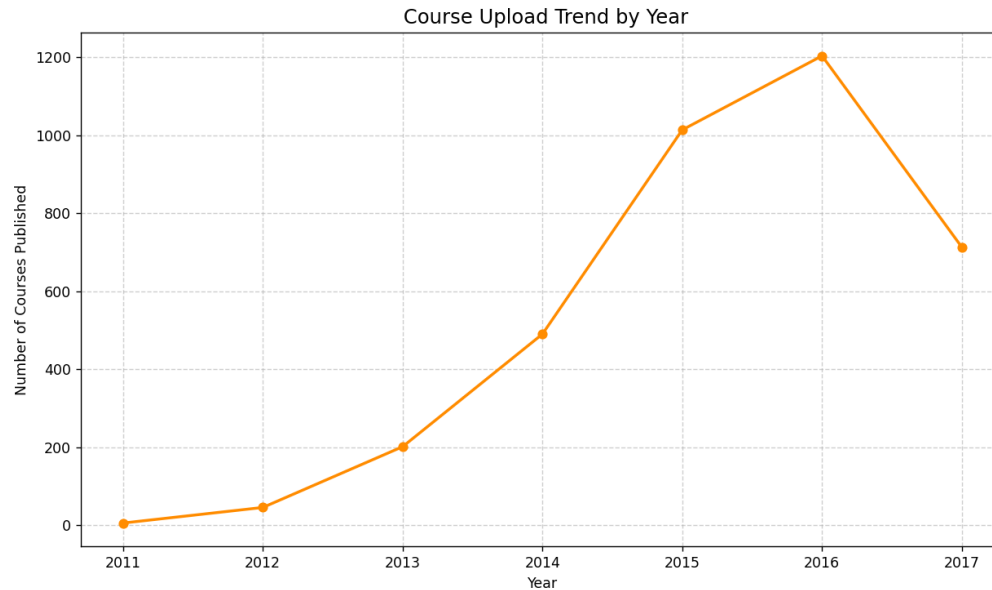
iii. Analysis Results

After preprocessing the dataset to convert and extract years from the `published_timestamp` column, a grouped count was performed to obtain the total number of courses uploaded each year. It was observed that:

- There was a notable **increase** in course uploads during certain years (e.g., peaks around 2016).
- A **decline** was noticed 2017, likely due to platform saturation or strategic shifts.
- The graph clearly shows a **significant surge** in 2016, followed by a **drop** in 2017 and fluctuations afterward.

iv. Visualization

For visual representation, a **line plot** was used to track the course uploads over time. The X-axis represents the **years**, while the Y-axis shows the **number of courses published** in that year.



Objective 6: Courses Published by Day of the Week

Analyze which days of the week instructors publish the most courses.

i. General Description

This objective aims to uncover the day-wise publishing pattern of courses on the platform. By analyzing which **days of the week** instructors most frequently publish their courses, we can detect behavioral patterns and preferences among content creators. This can also help platform managers optimize their backend processes and promotional strategies around peak publishing days.

ii. Specific Requirements

- Convert the `published_timestamp` to a proper datetime format.
- Extract the **day of the week** from each timestamp.
- Count the number of courses published on each weekday.
- Ensure correct ordering (Monday through Sunday).
- Visualize the results in a simple and intuitive bar chart.

iii. Analysis Results

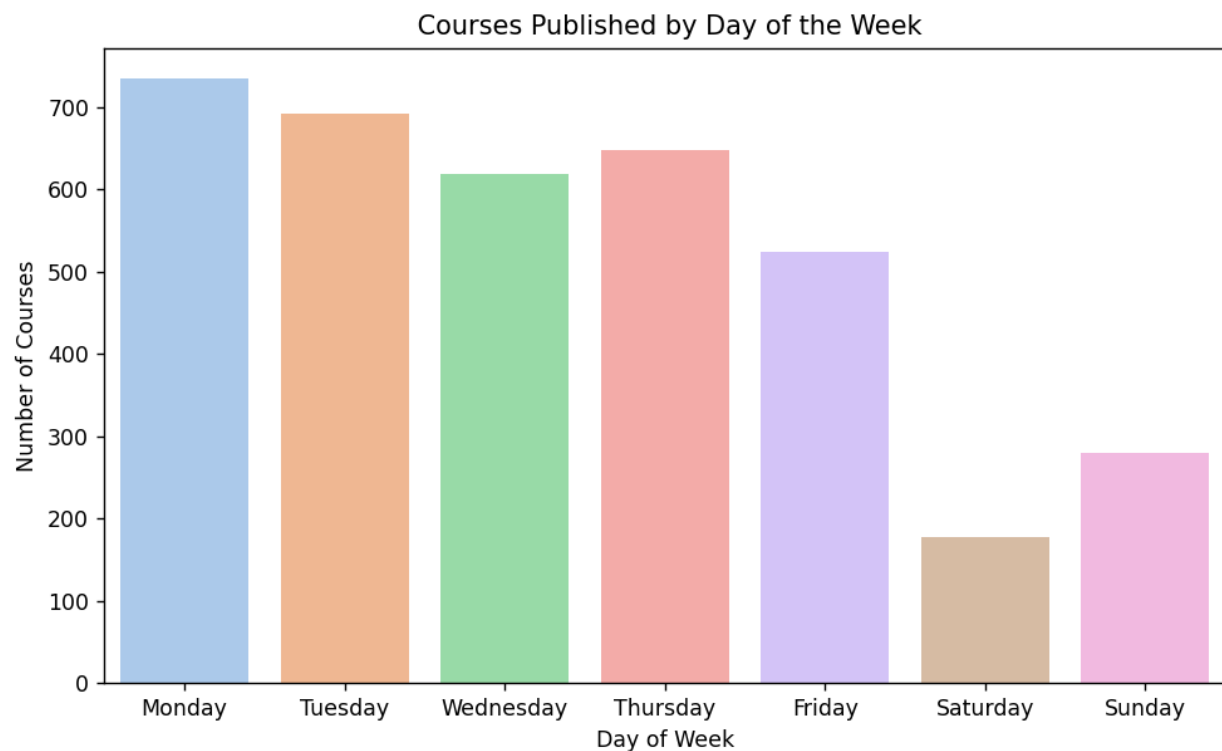
The analysis revealed interesting insights into instructor behavior:

- Most courses are published during **weekdays**, especially **Tuesdays** and **Wednesdays**.
- **Weekends** (Saturday and Sunday) see a noticeable **drop** in course uploads.
- This may suggest that instructors prefer to work on and finalize uploads during the week when engagement and platform activity might be higher.

This pattern can be useful for optimizing content management and even for learners looking to discover new courses earlier in the week.

iv. Visualization

A **bar plot** is ideal for clearly showing which weekdays dominate course publication. The bars are sorted in logical weekday order from Monday to Sunday.



CONCLUSION

This project aimed to explore and uncover insights from the Udemy course dataset by performing thorough data preprocessing, cleaning, and exploratory data analysis (EDA).

Through the lens of six well-defined and creative objectives, we examined various aspects of course content, popularity, publishing trends, and instructor behavior.

From the comparison of free vs. paid courses across subjects, we discovered how different content types vary in accessibility. The analysis of course duration revealed patterns indicating the preferred length for higher student engagement. Additionally, identifying the most popular course titles helped highlight key areas of learner interest.

By visualizing review distributions across course difficulty levels, we gained clarity on which skill levels drive user feedback the most. The study of course creation trends over time unveiled how course publishing activity has evolved, influenced possibly by industry and global events. Lastly, analyzing publishing patterns by day of the week offered behavioral insights into instructor activity.

These analyses not only demonstrated meaningful patterns and correlations but also showcased the value of data-driven decision-making in online education platforms. The project effectively combines technical analysis with storytelling, helping convert raw data into actionable intelligence. Future extensions could involve predictive modeling or text-based analysis on course descriptions for even deeper insights.

FUTURE SCOPE

While this project provided valuable insights into Udemy's course data, there remains significant potential for deeper and broader analysis in future work. Some possible directions for future enhancement include:

1. **Predictive Modeling:** Machine learning models could be developed to predict the popularity of a course based on its attributes such as title, subject, level, and duration. This would help instructors optimize their content creation strategies.
2. **Sentiment Analysis on Reviews:** If access to review texts is obtained, natural language processing (NLP) techniques can be used to analyze sentiments and understand learner feedback on a more granular level.

3. **Time Series Analysis:** A deeper investigation of course publishing trends over months and years could uncover seasonal patterns or the impact of external events (e.g., the COVID-19 pandemic) on online education.
4. **Geographic and Demographic Analysis:** Incorporating data on where students come from and their backgrounds can provide personalized learning insights and help platforms cater to a global audience more effectively.
5. **Pricing Strategy Optimization:** With more detailed pricing and enrollment data, advanced statistical analysis could be used to determine optimal pricing strategies for different course categories.
6. **User Behaviour Tracking:** Integrating data related to course completion rates, dropout rates, and user navigation behaviour could provide a more holistic understanding of learner engagement and areas for improvement.


By expanding the dataset or combining it with other relevant sources, future analyses can become more insightful and impactful, ultimately contributing to better educational outcomes and platform development.

REFERENCES

- [1] "Udemy Courses Dataset," Github, [Online]. Available: [Datasets/winequality-white.csv at master · MainakRepositor/Datasets](#) [Accessed: Apr. 4, 2025].
- [2] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed. Sebastopol, CA: O'Reilly Media, 2017.
- [3] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*, Sebastopol, CA: O'Reilly Media, 2016.
- [4] M. L. Waskom, "Seaborn: Statistical Data Visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.
- [5] "Matplotlib — Visualization with Python," The Matplotlib Development Team, [Online]. Available: <https://matplotlib.org>. [Accessed: Apr. 10, 2025].
- [6] "Pandas Documentation," The Pandas Development Team, [Online]. Available: <https://pandas.pydata.org/docs/>. [Accessed: Apr. 10, 2025].

SCREENSHOTS:

LinkedIn Screenshots:



Bhoomi Yadav • You
Aspiring Data science engineer | Python Enthusiast | Experience in python, Jav...
1d • 🌐

Just wrapped up an insightful Data Science project!

Project Title: Exploratory Data Analysis on Udemy Courses Dataset

This project explores patterns and trends in online education using real-world Udemy course data.

Here are the key objectives I focused on:

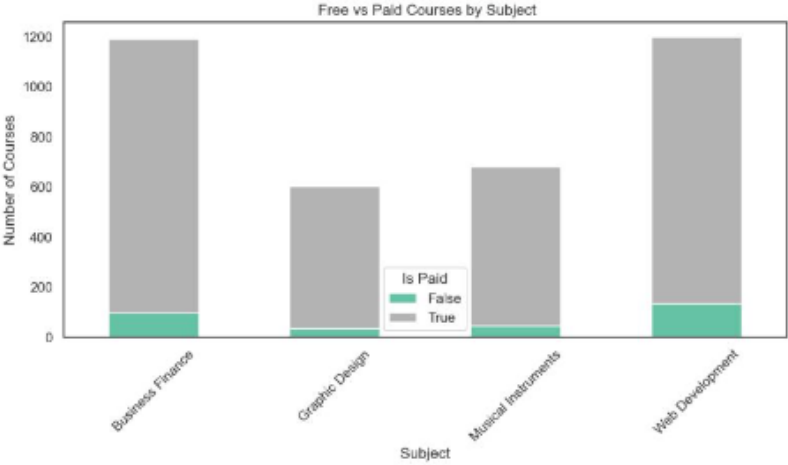
- Free vs. Paid Courses by Subject – Compared how many free and paid courses are offered across different subject areas.
- Optimal Course Duration – Identified the ideal course length that attracts the highest number of students.
- Most Popular Course Titles – Highlighted course titles with the highest enrollments.
- Course Reviews by Difficulty Level – Analyzed which course levels receive the most reviews.
- Trends in Course Creation Over Time – Tracked how course uploads have evolved year by year.
- Courses Published by Day of the Week – Analyzed which weekdays see the most course publications.

Tools used: Python, Pandas, Seaborn, Matplotlib, NumPy




From data cleaning to visualization, this project was a deep dive into the world of online learning!


Proud to turn data into decisions!

[#DataScience](#) [#Python](#) [#EDA](#) [#UdemyDataset](#) [#Kaggle](#) [#DataAnalytics](#) [#Seaborn](#) [#Matplotlib](#) [#StudentProject](#) [#LinkedInProjects](#) [#ExploratoryDataAnalysis](#)



Subject	Free (False)	Paid (True)
Business Finance	~100	~1100
Graphic Design	~50	~550
Musical Instruments	~50	~600
Web Development	~150	~1050





You and 53 others

+2

25 comments

Reactions

GitHub Screenshots:

The screenshot shows a GitHub repository page for 'Analysis_on_udemy_courses' by user Bhoomi64. The repository is public and has 5 stars, 1 watch, and 0 forks. The main branch is 'main' with 1 branch and 0 tags. The repository contains two files: 'README.md' (updated 18 hours ago) and 'main.py' (updated 22 minutes ago). The commit history shows 12 commits by Bhoomi64, with the latest commit 'Update main.py' made 22 minutes ago. The README file is titled 'Analysis_on_udemy_courses' and describes the analysis of Udemy courses using Python. The repository also has a 'Releases' section with no published releases and a 'Packages' section with no published packages. The 'Languages' section shows that the repository is 100.0% Python.

Repository: Analysis_on_udemy_courses (Public)

Branches: main (1 Branch), 0 Tags

Commits: 12 Commits

Files:

File	Update	Time
README.md	Update README.md	18 hours ago
main.py	Update main.py	22 minutes ago

README

Analysis_on_udemy_courses

analysis on Udemy courses using python language objectives:-

1. Identify course titles that attracted the highest number of students.
2. Compare how many free and paid courses are offered in each subject area.
3. Find the ideal course length that attracts the most students.
4. Compare the average duration of free and paid courses to see if paid ones offer more content.
5. Analyze which course level (Beginner, Intermediate, etc.) gets the most reviews.
6. Track how course uploads have changed over the years.

About

No description, website, or topics provided.

Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Languages

Python 100.0%