An Exploratory Analysis on Video Stats Dataset

Jimmy Cabrera - 230702295

2024-11-13

Introduction

This report provides an analysis of data from seven years of an open online course (MOOC) developed by Newcastle University and delivered through the online platform 'FutureLearn'. The course, titled "Cyber Security: Safety At Home, Online, and in Life", was a three-week program, free for anyone to access.

The data includes metrics collected by FutureLearn, tracking learner engagement and interaction throughout the course. In this report, a specific research question related to learner engagement in video content is investigated, using exploratory data analysis aligned with the Cross Industry Standard Process for Data Mining (CRISP-DM). This report will provide insights into video engagement trends and other relevant statistics across the seven-year span.

Round 1 of the CRISP-DM Cycle

Business Understanding

The initial phase of this analysis process involves firstly establishing the business objectives, identifying relevant stakeholders, and outlining clear aims for the desired outcomes. This stage also entails defining success criteria to assess the effectiveness of the results. Based on the insights gathered during this phase, a detailed plan has been created, to align with the principles and structure of the CRISP-DM framework. This phase of the CRISP-DM cycle is essential for ensuring that the analysis is purposeful, targeted, and relevant to the stakeholders involved.

Objective

The primary objective of this analysis is to provide valuable insights into the video engagement data for the course "Cyber Security: Safety At Home, Online, and in Life". By analyzing key metrics, such as video view counts, device usage patterns, and regional engagement statistics, this research aims to uncover trends and patterns that can be leveraged to improve course content and learner engagement. The analysis is specifically intended to help course designers and educators better understand how learners interact with the course videos, which video content resonates most with the audience, and where potential improvements can be made to enhance the overall learning experience.

Stakeholders

The key stakeholders in this analysis are the course developers, educators, and administrators at Newcastle University, as well as the online learning platform FutureLearn. These stakeholders will benefit from the findings as they can use the insights to make data driven decisions regarding course content, structure, and marketing strategies. By understanding which video segments capture learners' attention and which areas might need adjustment. From here the stakeholders can optimize the course to increase engagement and completion rates. Additionally, this information can be useful to instructional designers aiming to enhance learner outcomes by aligning content with the preferences and needs of the target audience specially when dealing with other regions where access could be limited.

Analysis Goals

The specific goals of this analysis are as followed:

- To identify patterns in video engagement, including total views, downloads, and caption/transcript views, across different video segments.
- To analyze user behavior related to video completion rates at various viewing thresholds (e.g., 5%, 50%, 100%) and device usage trends (e.g., mobile, desktop, tablet).
- To explore geographical variations in engagement, focusing on how learners from different continents interact with the course videos.

Success Criteria

Success in this analysis will be determined by the ability to provide actionable insights that address the stated goals. The data should be presented in a way that is easily interpretable by stakeholders. The final analysis will offer clear recommendations for improving the course's content, delivery, and overall user experience, which can then be used to inform decisions regarding future iterations of the course. The success of the investigation will also be evaluated based on how effectively it helps stakeholders understand the key drivers of engagement and how these insights can be leveraged to enhance the learning experience.

Initial Research Question

In alignment with the objective of this report, the initial research question guiding this analysis is:

"How does learner engagement with the video content vary across different segments of the course, and how do factors such as device type, geographic location, and viewing patterns impact overall video performance?"

By addressing this question, the analysis will provide a detailed exploration of video engagement, offering insights that are both specific and actionable for stakeholders seeking to improve the course's effectiveness.

Data Understanding

In Phase 2 of the CRISP-DM cycle, the focus shifts to a much thorough understanding of the data. This involves assessing the data requirements based on the objectives outlined in the previous phase, evaluating the availability of relevant data, and determining the reliability of the data sources used. This step is crucial to ensure that the data gathered aligns with the business objectives and supports the overall goal of the analysis. A careful review of the collected data is essential to confirm its suitability for the analysis and to refine the approach if necessary.

Data Collection

The data used in this analysis was sourced from the FutureLearn platform, which contains detailed engagement statistics for the course "Cyber Security: Safety At Home, Online, and in Life." The data set comprises multiple variables related to learners' interactions with the course videos, collected over the course of seven years. The raw data contains information on video statistics, such as total views, downloads, device usage, regional distribution, and video completion rates. This data, which is structured into specific rows corresponding to individual course video segments, provides a good source of insights into learner behavior and video performance.

Exploring the Data

Upon gathering the dataset, the next step was to explore its contents in detail. This exploration involved reviewing the structure of the data, identifying any potential issues with data quality, investigating the variable types, and assessing whether the available data aligned with the research objectives.

The dataset contains several key variables related to video engagement, including:

• Video Duration • Total Views • Total Downloads • Total Caption Views • Total Transcript Views • Completion Rates (e.g., views at 5%, 50%, 100%) • Device Usage (e.g., mobile, desktop, tablet, TV) • Geographic Distribution (views by region)

After reviewing the dataset, it was evident that it contained sufficient information to address the research question regarding learner engagement with video content. The data was collected consistently across all years, with clear variables representing learner interactions, such as video views, device types, and geographic regions. However, there were some nuances to consider, such as the lack of specific units of measurement for certain variables. This issue could complicate the interpretation of certain data points, but it was determined that this would not significantly affect the overall analysis, as the relationships between the key variables were still discernible.

The next step in the exploration was to ensure that any gaps in the dataset, such as missing or incomplete data, would not impede the analysis. In this case, the dataset was found to be largely complete, with no major data quality issues identified.

Summary of Findings

The key variables necessary for answering the research question were present, including video engagement metrics, device usage statistics, and geographic distribution of views. Although there were minor issues regarding the labeling of units and variable formats, these did not significantly hinder the ability to carry out the analysis. It was also noted that the dataset contained enough data points to meet the objectives of the research without the need to adjust the scope. The data was deemed reliable, with no major errors or discrepancies. Overall, the dataset was found to be well-suited for analysis, and the next step would involve preparing the data for further investigation in line with the objectives outlined in the business understanding phase.

Data Preparation

Data preparation, involves cleansing, transforming, and selecting data in preparation for the modeling phase. The importance of this phase is to ensure that the dataset is in a consistent, tidy format that facilitates analysis and reduces the risk of errors. By normalizing the data and addressing any inconsistencies, this step ensures that the dataset is ready for detailed analysis, making the process more efficient and ensuring that future users of the data can confidently reproduce the results.

Data Cleansing

The first step in the data preparation process involved categorizing each set of data. I identified the full set of variables present across the two years and compiled a list of variable names. These names were then applied to each variable in both years to standardize the dataset. This step helped ensure that the variables were easily interpretable and reduced potential confusion when working with the data, as well as separating specific data from both years with ease. All the data gathered from these two years (as we are just focusing on two years worth of data) where grouped into one group, and then separated into compartments in order to easily access each specific row with its own set of readings. The reason for this is to calculate each set quicker, and create a variety of ranges in order to conceive the correct data representation measurements, and expel results that would benefit our main research question.

Data Wrangling

Following the data cleansing, I proceeded with the data wrangling phase, transforming and deriving variables from existing data to create a comprehensive dataset that maximized the available information for both years. The transformation process enabled me to ensure that the two datasets were aligned in terms of variables, facilitating direct comparisons between the two years.

This approach also improves the efficiency of code construction, as it allows for easy adaptation should additional years of data be added or if further automation is needed. The organization of the code and use of helper functions enhances its readability, making the analysis more accessible for anyone who may need to reproduce or evaluate the process in the future.

After deriving the necessary variables, I ordered the columns into a common sequence, allowing for easier comparisons between the two years. This step ensured that the structure of the dataset remained consistent across both years, simplifying subsequent analysis.

Data Combination

The next step in the wrangling process involved combining the two years' data into a single, unified dataset. This task was done by using helper functions to identify any missing headers between the two datasets and adding these variables with 'NA' values where applicable. A new variable was created in each table to denote the year of the observation, ensuring that the data could later be split or analyzed by videos as needed. Finally, the columns were once again arranged in a standardized order, and all data from both years was appended into a single table.

Analysis of the First Two sets of data

This analysis involved separating each row of data (representing different video segments) and calculating the mean of each variable across all observations for both years. By doing so, I was able to obtain a consolidated view of the video engagement statistics, providing a baseline from which comparisons could later be made with any other data input.

The focus of this initial analysis was to explore the patterns of learner engagement with the course's video content. By calculating the mean values for each video segment, I was able to determine the average performance across various metrics, such as total views, downloads, completion rates at various thresholds (e.g., 5%, 50%, 100%), and device usage percentages (e.g., mobile, desktop, tablet). This provided an overview of how learners interacted with the course videos during the first two years of its availability.

Additionally, by comparing these metrics across the different video segments, I could identify which videos performed particularly well in terms of engagement and which videos might benefit from improvements or adjustments in future iterations of the course.

Modelling

The focus now shifts to applying appropriate modeling techniques to analyze the data and derive insights that can answer the research questions posed in the business objective. This phase leverages the structured and cleaned data from the previous steps, utilizing statistical and machine learning models to identify patterns, relationships, and trends. For this analysis, the data from the first two years were used to explore learner engagement with video content across different segments of the course, considering factors such as device type, geographic location, and engagement patterns. Various visualizations were created to model these relationships and gain insights into how these factors impact overall video performance.

The following models were used to represent the data and explore key aspects of learner engagement:

Bar Plot for Device Type Usage Percentage

The chart presents the percentage of usage for each device category 'Console, Desktop, Mobile, TV, Tablet, and Unknown' across the two years. The results indicate a clear preference for desktop devices, with a significant proportion of learners accessing the course materials through desktops and smartphones. This aligns with the growing trend of mobile learning, reflecting learners' preference for flexible access to educational content. The relatively lower percentages suggest that these devices are less commonly used for online learning activities.

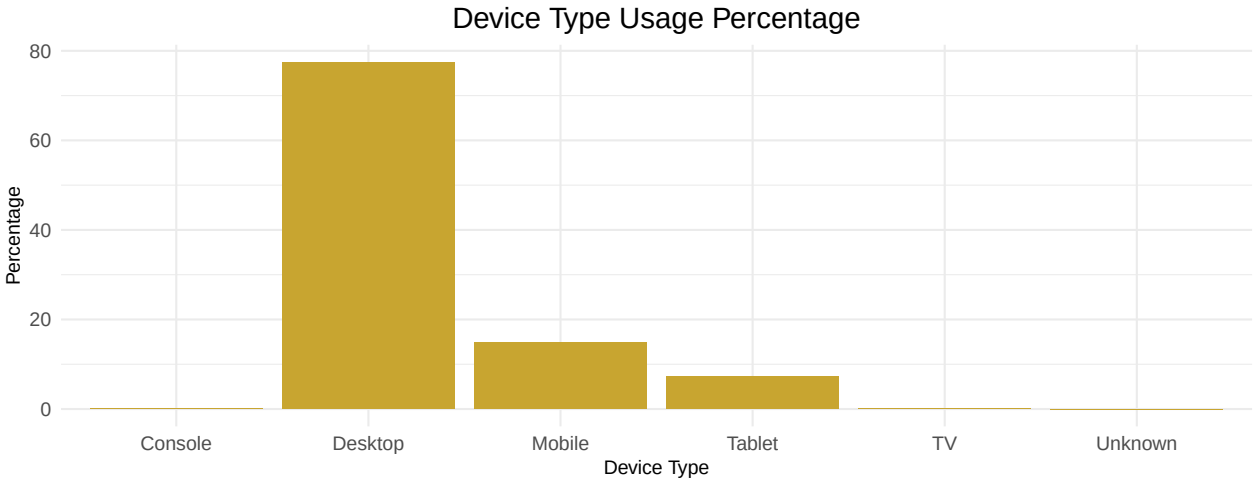
Interpretation: The bar plot highlights the increasing reliance on desktop but moving gradually to mobile devices for online learning, signaling a shift towards mobile first learning experiences. This finding suggests that course content should be optimized for mobile users, ensuring that it is easily accessible and engaging on smaller screens.

Overall Metrics for Two Years

Metric1	Metric2	Metric3	Metric4
step position	viewed hd	viewed onehundred percent	europe views percentage
title	viewed five percent	console device percentage	oceania views percentage
video duration	viewed ten percent	desktop device percentage	asia views percentage
total views	viewed twentyfive percent	mobile device percentage	north america views percentage
total downloads	viewed fifty percent	tv device percentage	south america views percentage
total caption views	viewed seventyfive percent	tablet device percentage	africa views percentage
total transcript views	viewed ninetyfive percent	unknown device percentage	antarctica views percentage

Overall Values for Two Years

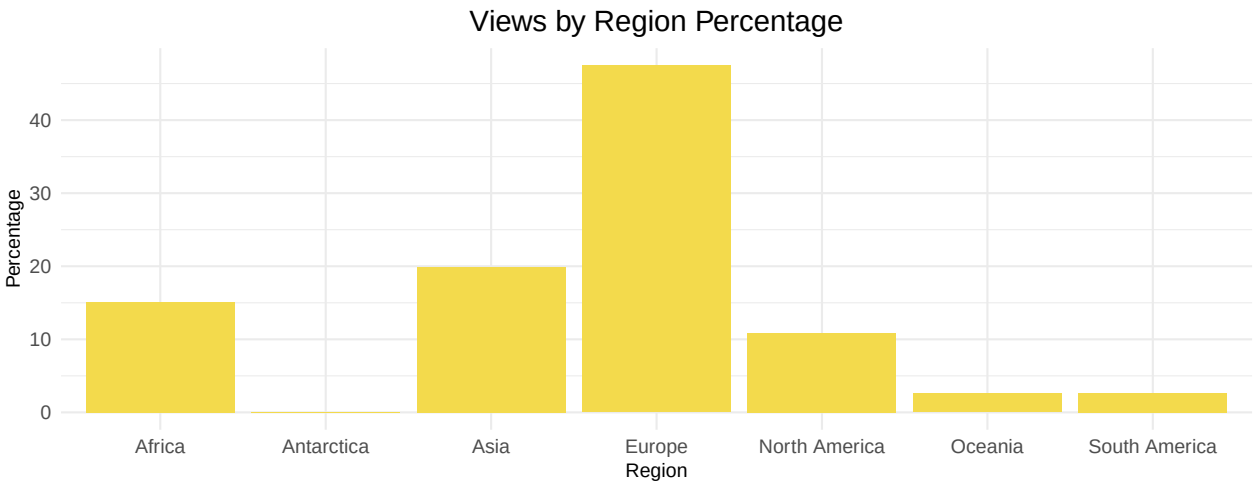
Value1	Value2	Value3	Value4
1.1	59.50	63.66	47.47
NaN	78.06	0.03	2.60
99.0	76.80	77.47	19.92
1774.5	74.46	14.89	10.87
152.5	70.86	0.03	2.64
37.5	67.94	7.24	15.13
243.5	66.21	0.00	0.00



Bar Plot for Regional Distribution of Learner Engagement

The graph reveals that Europe and Asia are the largest regions contributing to overall engagement, followed by Africa and North America. The smaller percentages for regions such as Oceania, South America, and Antarctica suggest that these areas have a lower proportion of learners participating in the course.

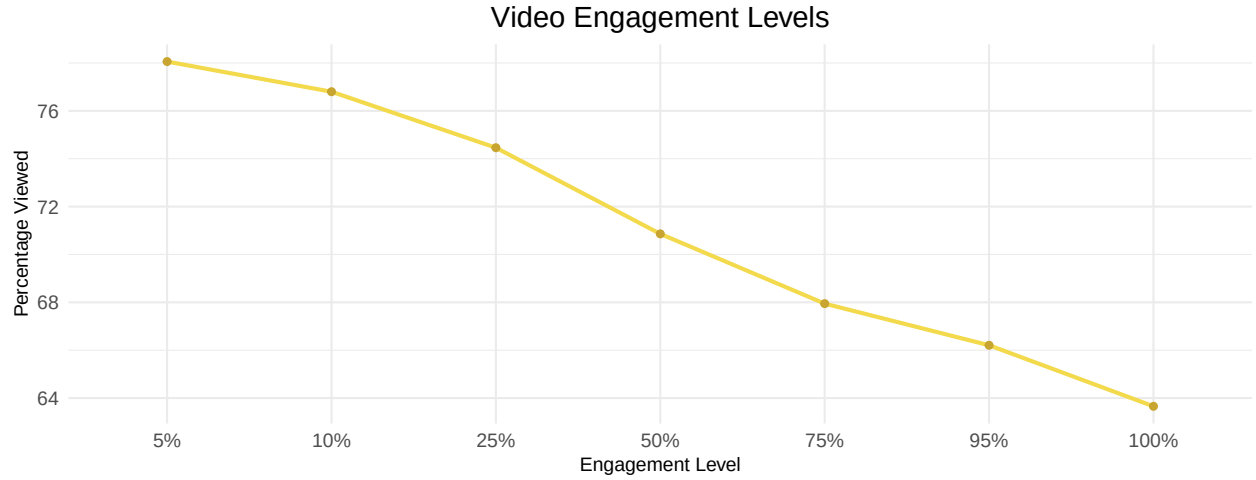
Interpretation: The regional bar plot provides insight into the global reach of the course, indicating that the course content has wider engagement in certain regions, particularly in Europe and Asia. This geographic trend may influence future decisions regarding localized content, targeted marketing, or regional adjustments in course delivery to improve engagement in less represented regions.



Line Plot for Engagement Metrics Across Different Levels

The chart shows the percentage of viewers who watched varying amounts of each video, from 5% up to 100%. The results suggest that there is a noticeable drop in engagement after the initial 5% of the video, with significant declines as the video progresses. However, it does show that the 100% completion rates still achieve a reasonable percentage of views.

Interpretation: The line plot reveals typical video consumption patterns, with a large initial drop-off in viewer engagement. This suggests that learners tend to lose interest early in the video or that the content might be less engaging at the start. This insight calls for improvements in the introduction or opening segments of the course videos to retain learner attention and encourage deeper engagement throughout the video. Additionally, strategies such as adding interactive elements or breaking the content into smaller, more digestible segments may help maintain engagement across the video.

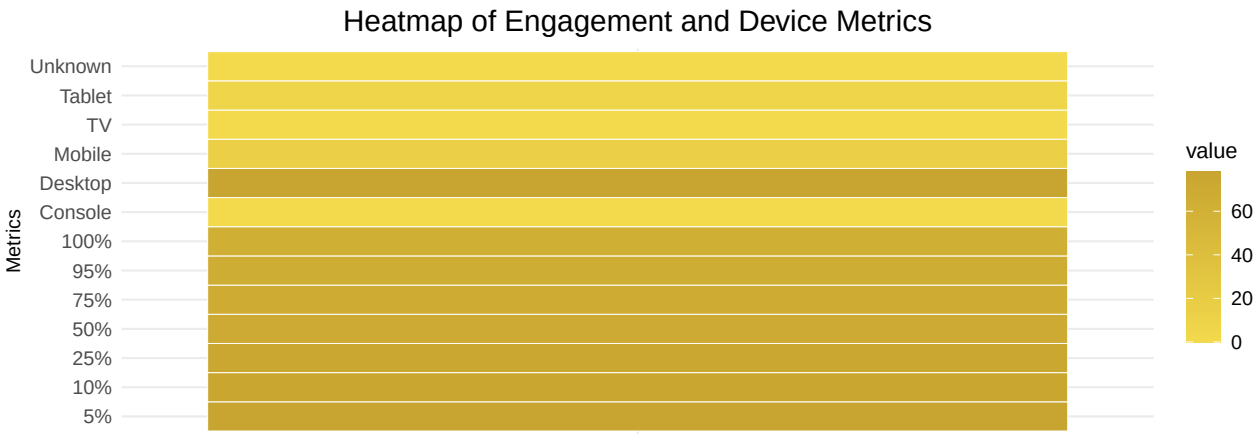


Heatmap for Engagement and Device Metrics

The data points represent the average percentage of engagement for each device type at various levels of video consumption (from 5% to 100%). The heatmap indicates that desktop devices show higher engagement percentages at the lower viewing levels (e.g., 5%, 10%, 25%) compared to other device types. However, there is a noticeable decrease in engagement as the viewing level increases, especially for mobile and tablet devices.

Interpretation: The heatmap provides a visual correlation between device usage and engagement levels. Desktop devices, while leading in early engagement, seem to experience higher drop-off rates as the video progresses. This suggests that while desktop devices are preferred for accessing content, engagement strategies should be refined to improve retention, particularly for mobile users. It may be worth exploring whether different forms of content, such as

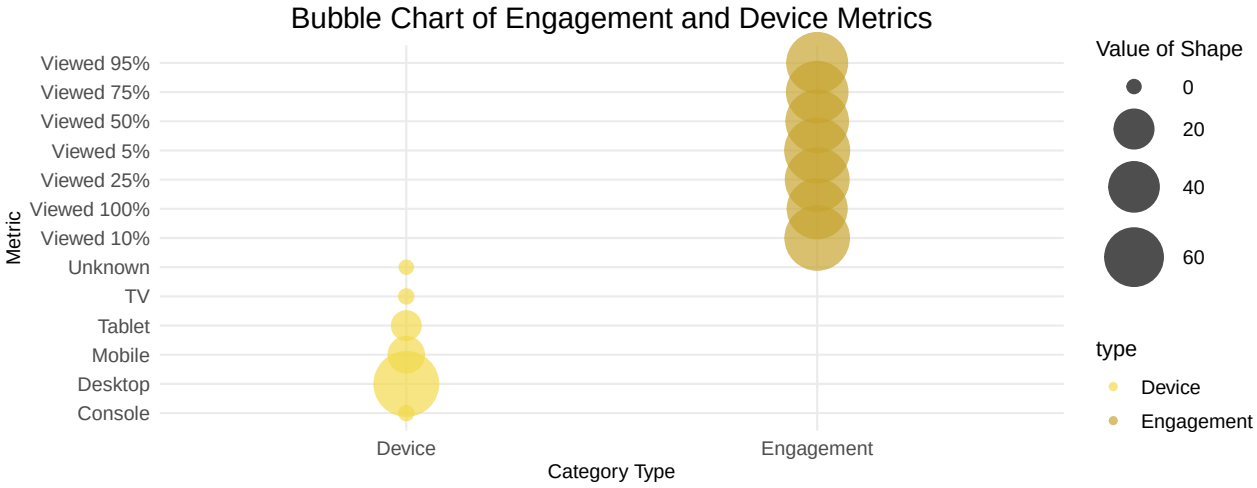
videos optimized for smaller screens or supplementary interactive materials, could enhance the viewing experience on mobile devices.



Bubble Chart for Combined Engagement and Device Metrics

The size of each bubble corresponds to the percentage value of either engagement or device usage, while the color differentiates between engagement levels and device types. This chart offers a clear comparison between how various devices impact the engagement levels and highlights the significant variance in engagement across devices.

Interpretation: The bubble chart illustrates the correlation between the device used and the engagement level, with desktop devices showing higher engagement at the lower levels of video consumption. The size of the bubbles allows for a quick comparison of the relative importance of each category. This visualization reinforces the need for mobile optimized content to cater to the majority of learners engaging through mobile devices.



Round 1 Evaluation

Phase 1 of CRISP-DM provided valuable insights into learner engagement across device types, geographic locations, and viewing behaviors. The structured data preparation enabled seamless analysis, while a variety of visualizations clarified key engagement patterns. For example, bar plots revealed a preference for desktop device access, regional distribution highlighted higher engagement in Europe and Asia, and line plots showed significant early video drop-offs. Additionally, heatmaps and bubble charts underscored device-specific engagement patterns, indicating that mobile users may need targeted retention strategies due to their higher initial engagement but quicker drop-off rates. These visualizations provided a foundational understanding of when and how engagement declined, suggesting the need for a

more compelling early course design to retain interest.

This analysis reveals that early video drop-offs and device-specific patterns are critical areas for optimizing course content. The findings underscore the importance of mobile optimization and engagement strategies, particularly in regions with high viewer numbers. Although the two-year dataset provides substantial insights, incorporating additional years and data on session duration or demographic segmentation would enhance the reliability and granularity of future evaluations. Moving forward, these results will guide adaptations in content pacing, device-specific engagement tactics, and regionally tailored outreach, thereby supporting a more effective, learner-centered course design.

Round 2 of the CRISP-DM Cycle

Business Understanding Review

Following the results from Phase 1, which provided foundational insights into learner engagement for MOOC, the objectives of this analysis continue to focus on understanding how learners interact with course video content. With no obstacles identified in Phase 1, the analysis now shifts to a longitudinal approach, investigating changes in learner engagement over a broader five year period, incorporating data from three additional years.

The refined research question for this second phase is as follows:

"How has learner engagement with video content evolved over the 5-year period, and what trends can be observed in device usage, geographic location, and completion rates across the years?"

This question is pertinent given the observable shifts in digital education and learner behavior over time. The analysis will offer stakeholders—course designers, educators, and platform administrators deeper insights into longitudinal engagement trends and device and regional variations. The goal is to identify persistent engagement patterns, areas for improvement, and targeted strategies to enhance learner retention and content relevance. Given that Phase 1 objectives and success criteria were met, this analysis will proceed with the existing scope and measures.

Data Understanding

This phase necessitates a fresh examination of the data to ensure its suitability for answering the new research question. Given the focus on trends across three additional years, any data limitations or structural changes need addressing to ensure consistency. The initial analysis confirmed the adequacy and reliability of the core dataset from FutureLearn, and the same source will support this expanded analysis.

The data set now spans five years and includes variables crucial for understanding engagement trends, such as video views, device usage, geographic location, and completion rates. As in Phase 1, the data quality is consistent, though some adjustments, such as standardizing variable labels across years, may be required for coherence.

The completeness and reliability of the data suggest that it is well-suited for this expanded analysis, with minimal adjustments needed. The data will be prepared for Phase 2's deeper analysis by deriving new variables that encapsulate the multi year trends.

Data Preparation

The data from each year is cleansed and formatted to maintain consistency across variables, facilitating reliable year over year comparisons, similar t phase 1 preparation.

Data Standardization: All variable names and data structures were standardized across all years, ensuring that each dataset conformed to a uniform structure.

Derived Variables: New variables were created to capture year over year trends, including multi-year averages and year-specific summaries for each variable.

Data Binning: To simplify trend analysis, continuous variables like completion rates and total views were binned into categories based on predefined thresholds. This binning enables clearer visual comparisons and insights into engagement patterns.

Overall Metrics for All 5 Years

Metric1	Metric2	Metric3	Metric4
step position	viewed hd	viewed onehundred percent	europe views percentage
title	viewed five percent	console device percentage	oceania views percentage
video duration	viewed ten percent	desktop device percentage	asia views percentage
total views	viewed twentyfive percent	mobile device percentage	north america views percentage
total downloads	viewed fifty percent	tv device percentage	south america views percentage
total caption views	viewed seventyfive percent	tablet device percentage	africa views percentage
total transcript views	viewed ninetyfive percent	unknown device percentage	antarctica views percentage

Overall Values for All 5 Years

Value1	Value2	Value3	Value4
1.1	50.60	64.66	45.52
NaN	81.22	0.01	3.08
99.0	79.42	73.60	25.87
1508.0	76.89	19.46	10.05
114.2	72.62	0.01	2.61
34.2	69.47	6.59	11.84
222.0	67.76	0.00	0.00

Consolidation and Grouping: The data from each year was combined into a unified dataset with an identifier for each year, facilitating longitudinal analysis. Helper functions were applied to manage missing headers across years, ensuring uniform data structure.

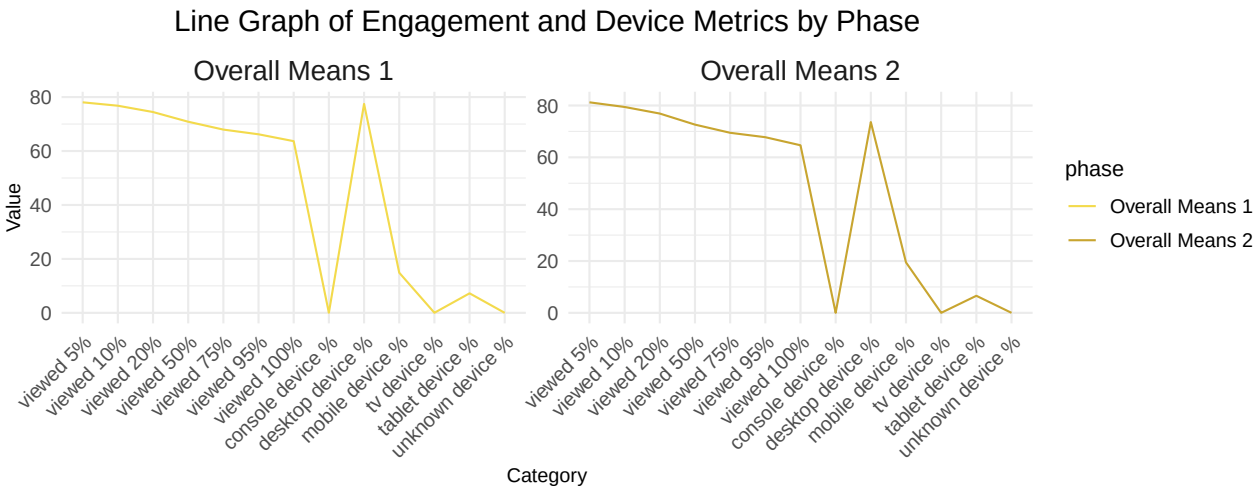
These steps resulted in a structured dataset ready for modeling and longitudinal analysis, providing a clear foundation for evaluating changes in engagement trends over the years.

Modelling

In Phase 2, the analysis shifts to modeling learner engagement across five years, examining how video interaction, device usage, geographic location, and completion rates have evolved. The modeling process centers on identifying trends and relationships that emerge over time, as well as pinpointing any significant variations in how learners interact with the course content.

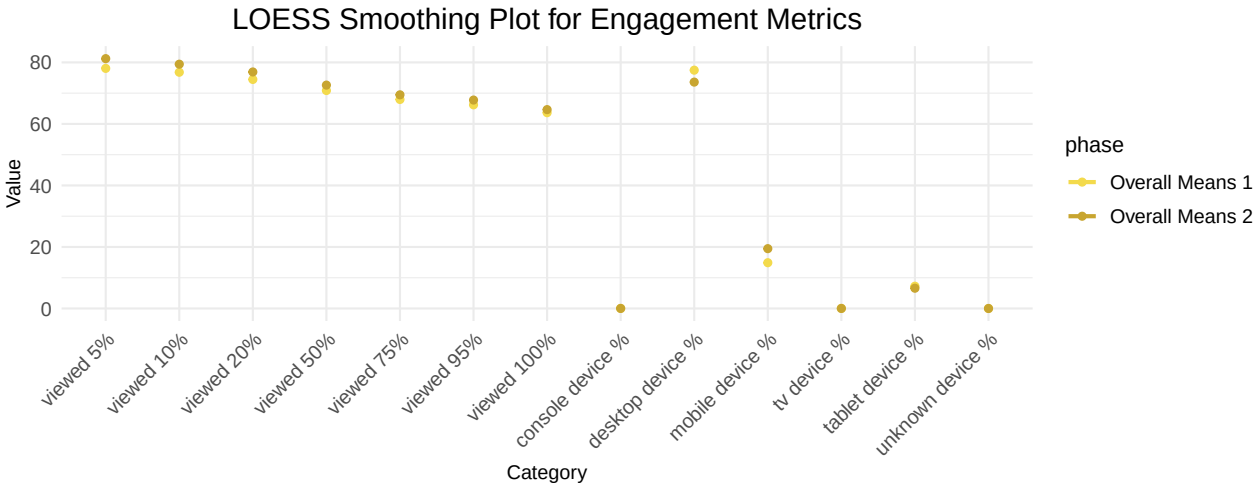
Trends in Learner Engagement

To capture changes in engagement over the five-year period, several visualizations were employed, each providing insight into different facets of learner behavior. The line graph with facets is particularly useful for understanding shifts in engagement metrics, such as video views and device usage, over time. By plotting these metrics separately for each year, we can observe distinct patterns across categories like “Viewed 5%” to “Viewed 100%” and device preferences (e.g., mobile, desktop, tablet). For example, the line graph reveals whether there has been a consistent increase or decrease in learner interaction with the course videos, particularly in the percentage of views completed and the devices used for accessing the content.



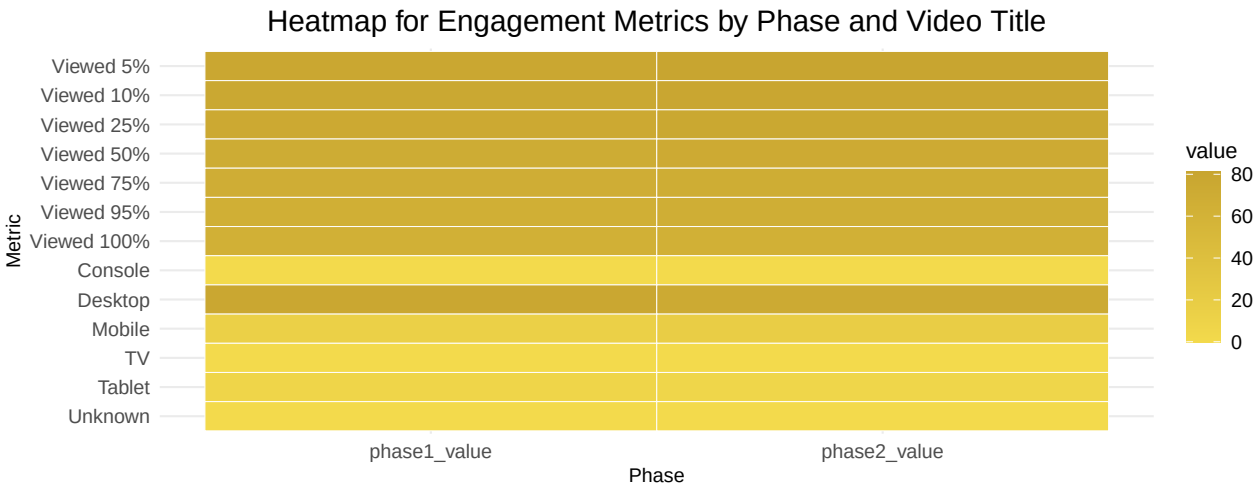
Moreover, the LOESS smoothing plot complements the line graph by providing a smoothed trend line that visualizes the overall engagement patterns across categories. This plot allows us to see how engagement levels evolve, accounting for any fluctuations while helping to identify long-term trends. The smoothed curves show how certain categories (e.g., mobile device usage or completion rates) have experienced notable growth or decline, highlighting shifts in learner behavior over time.

These visualizations provide clear evidence of engagement trends and device usage patterns across the five-year period, confirming the longitudinal shifts in how learners are interacting with the course content. The combination of line graphs and LOESS smoothing offers a comprehensive understanding of engagement dynamics.



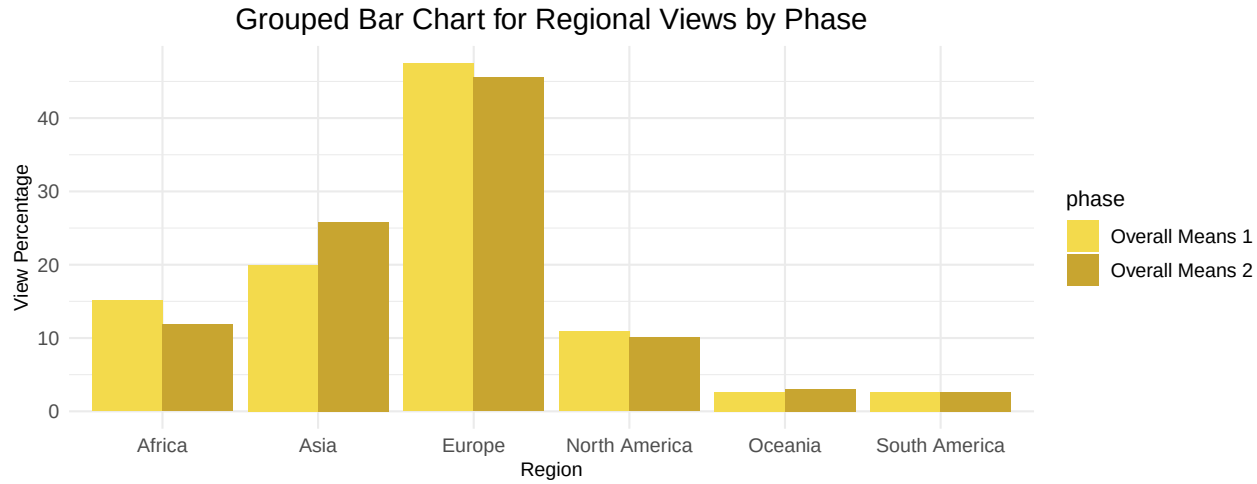
Device Usage and Geographic Distribution

The heatmap and grouped bar chart provide insights into device preferences and regional variations in learner engagement. The heatmap compares metrics across phases, displaying differences in engagement across video categories and phases. By observing the color gradients in the heatmap, we can easily identify which devices have seen a rise in usage over the years, as well as which metrics (e.g., “Viewed 50%” or “Viewed 100%”) show the most significant changes. This is especially valuable for understanding how different types of devices (desktop, mobile, tablet, etc.) contribute to overall engagement.



The grouped bar chart further illustrates regional engagement differences, showing how engagement metrics vary by region (Europe, North America, Asia, etc.). By comparing engagement rates across regions for each phase, we gain insight into how learners from different parts of the world engage with the content and how regional trends evolve. For example, we can analyze whether certain regions show an increase in mobile device usage, or if geographic shifts in learner behavior correlate with completion rates.

Together, these visualizations suggest that device preferences have changed over time, with a noticeable rise in mobile device usage, particularly in the later phases. Similarly, regional trends indicate that learners from certain regions may have increasingly relied on specific devices or demonstrated higher engagement levels in certain years.



Synthesis and Insights

The combination of these models highlights key trends in learner engagement, device usage, and regional differences:

Overall Engagement Trends: Over the five years, a gradual increase in video completion rates and device usage can be observed. This trend may reflect improvements in course design or changes in learner preferences.

Device Usage: Mobile devices have become changing over the years, with learners progressively shifting from desktop to mobile platforms. This trend is confirmed by both the line graph and heatmap, with mobile device usage showing a marked rise, particularly for higher completion categories (e.g., "Viewed 100%").

Regional Differences: The grouped bar chart reveals that regions such as Asia and Europe show higher engagement levels across phases compared to regions like South America and Oceania. These findings may suggest that access to devices or internet infrastructure impacts engagement, highlighting areas where targeted interventions could improve learner retention.

Round 2 Evaluation

The second phase of the CRISP-DM process has successfully addressed the research question regarding how learner engagement with video content has evolved over a five-year period. The analysis revealed significant shifts in device usage, with a notable increase in mobile device access, and highlighted regional differences in engagement patterns. These insights are crucial for course designers and platform administrators, as they provide actionable information for optimizing course content and delivery across various devices and geographies. Additionally, the study identified engagement behaviors that correlate with higher completion rates, offering valuable guidance for improving learner retention and course relevance.

The findings were presented in clear and accessible visualizations, including line charts, heatmaps, and other diagrams, to allow stakeholders to easily interpret the data and derive actionable insights. The use of reliable data from FutureLearn and the structured, longitudinal analysis ensured that the results were both reliable and relevant. Overall, this phase has successfully met the objectives of providing a comprehensive understanding of how learner engagement has evolved, and the insights generated can inform strategies to improve learner experience and retention in MOOCs moving forward.