BAN 693

CAPSTONE PROJECT

Udemy, LinkedIn and YouTube in the Landscape of Online Learning

Apurva Vaze	**Neenu Jose**	**Ruchika Vasudev Balani**	**Swetha Srinivasan**
os6218	**wg8684**	**ac9968**	**cx2194**

Table of Contents

Abstract

In the rapidly evolving landscape of online education, Udemy stands as a cornerstone platform, providing a diverse range of courses since its establishment in 2010. However, the challenge faced by instructors extends beyond delivering quality content to the effective expansion of their reach and engagement with their audience. Notably, platforms like LinkedIn and YouTube have emerged as influential players, reshaping the dynamics of online education.

Our study delves into this dynamic by closely collaborating with John Doe, a seasoned Udemy instructor grappling with the disparity between the exceptional quality of his courses and the anticipated audience engagement. Leveraging the transformative evolution of LinkedIn into a content-sharing platform and YouTube's strategic shift towards enabling content creators to monetize, our project investigates the correlation between social media engagement on LinkedIn and YouTube and the popularity of Udemy courses.

The primary focus of our analysis involves a comprehensive examination of Udemy, LinkedIn, and YouTube datasets. By scrutinizing variables such as Instructor Name, Number of students, Udemy Rating, LinkedIn and YouTube Followers, Posting Frequency, and engagement metrics, including likes, dislikes, views, and comments, we aim to provide empirical substantiation for the impact of social media engagement on Udemy course sales.

While our project acknowledges the importance of content engagement alignment, it intentionally narrows its scope to the specific dimension of engagement that has the potential to transform instructors' online presence. Through this research, we aspire to offer John Doe and other instructors valuable insights into optimizing their engagement strategies on LinkedIn and YouTube to enhance the popularity of their Udemy courses.

Introduction

Udemy, established in 2010 as an online learning platform, has since grown into a prominent hub where instructors offer an extensive array of courses. Instructors today are challenged to not only deliver high-quality content but also to expand their reach and engage their audience effectively. As a response to this challenge, platforms like LinkedIn and YouTube have gained remarkable significance in shaping the online education landscape.

LinkedIn, initially recognized as a platform for professional networking, has undergone a transformative evolution. It has evolved into a content-sharing platform, allowing instructors to showcase their expertise, share insights, and directly engage with their followers. Similarly, YouTube, originating as a video-sharing platform in 2005, has strategically shifted its focus to enable content creators to monetize their videos, revolutionizing the way knowledge and entertainment is shared.

In our current study, we have the privilege of collaborating closely with John Doe, an accomplished Udemy instructor who has meticulously developed courses that he firmly believes exemplify exceptional quality. However, despite his conviction, his courses aren't garnering the audience he anticipated. John is seeking answers as to why the excellence of his content isn't translating into the expected number of audiences.

Project Focus and Analysis

Recognizing the power of influence and footprint that instructors can establish on platforms like LinkedIn and YouTube, this project aims to determine whether a direct connection exists between LinkedIn and YouTube engagement and Udemy course popularity. By conducting a thorough examination and leveraging data-derived observations, our objective is to furnish John with empirical substantiation, demonstrating the influence of social media engagement on Udemy course sales.

To achieve these objectives, this project will comprehensively analyze Udemy, LinkedIn, and YouTube datasets. The primary aim is to unveil the correlation between instructors' LinkedIn and YouTube activities and their popularity on Udemy. The analysis will focus on gauging the impact of self-promotion efforts as manifested on LinkedIn and YouTube, utilizing variables such as Instructor Name, Number of students, Udemy Rating, LinkedIn and YouTube Followers, Posting Frequency, and engagement metrics like likes, dislikes, views, and comments.

Content Engagement Alignment

It's important to note that while this project will not extensively explore the alignment of content engagement between platforms and Udemy courses, our emphasis remains on understanding the degree of engagement Udemy instructors generate on LinkedIn and YouTube. This limitation is purposeful, focusing our analysis on a specific dimension of engagement that holds the potential to transform instructors' online presence.

Related Work

It is widely known that social media platforms are among the top choices for advertisers to target their audience, specifically it is the second most popular choice for digital ads, accounting for 33% of all digital advertising campaigns (McLachlan, 2023). There are plenty of guides, articles, tutorials and even courses that address strategies to market and boost massive open online courses (MOOCs) on social media platforms (Bankoff, 2023; Chaudhry, 2023). Such material often recommends actionable steps that can be taken by course instructors in order to increase the reach, and subsequently, the sale of the course (Notermans, 2023).

About the courses themselves, empirical studies identify certain video characteristics and formatting suggestions that maximize student engagement (Guo et al., 2014; Yang et al., 2022). Student engagement is an important measure of MOOCs' success since it determines course completion. Studies have shown that learners who comment on social learning communities are more likely to complete the course than those who do not engage in discussions (Crane & Comley, 2021). Additionally, students who engage in self-regulating learning strategies such as goal-setting and self-evaluation were found to have much higher completion rates compared to those who did not (Akinyemi, 2023). It has also been established that increasing engagement with students by way of social media communities, Facebook Groups, Instagram Stories, comments and discussion forums are vital to student success in MOOCs as defined by course completion and achieving learning outcomes (*Faculty Voice: Enhancing The Course Experience Using Social Media | Digital Learning & Innovation*, n.d.). Along similar lines, it has been established that external social media platforms like Facebook Groups show higher student engagement and retention than internal discussion forums in MOOC platforms (Zheng et al., 2016).

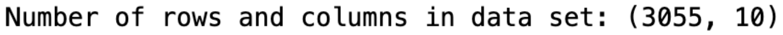
Existing literature discuss aspects of online learning such as student engagement, retention,

completion rates, video characteristics and such, while the research proposal is a focused plan to investigate the relationship between engagement on social media platforms and the popularity of Udemy courses. That remains the unique contribution of this paper.

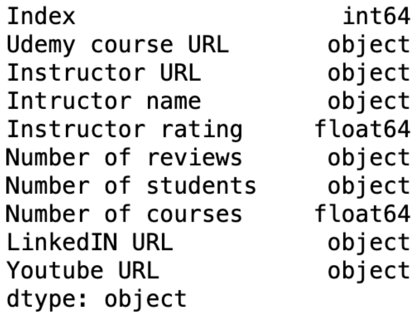
Dataset

Two distinct Excel files are utilized: one containing the top 5000 courses in the development category and the other featuring 3000 courses in the business category on Udemy. Both datasets were obtained from Kaggle and serve as the initial sources for acquiring Udemy course URLs. The data extraction process involves web scraping to obtain information from respective URLs. Specifically, social media URLs for platforms like YouTube and LinkedIn are acquired through the web scraping of Udemy URLs. After web scraping, there are four CSV files for each platform, in which two files are for LinkedIn - one for individual profiles and the other for profiles of organizations.

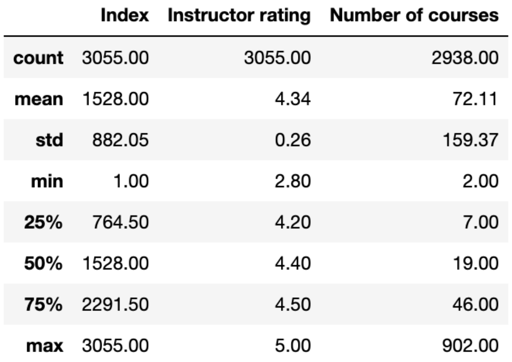
Udemy dataset obtained by web scraping from initial Kaggle dataset



```
Number of rows and columns in data set: (3055, 10)
```

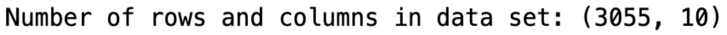


```
Index                    int64
Udemy course URL        object
Instructor URL          object
Intructor name          object
Instructor rating      float64
Number of reviews       object
Number of students      object
Number of courses      float64
LinkedIN URL            object
Youtube URL             object
dtype: object
```

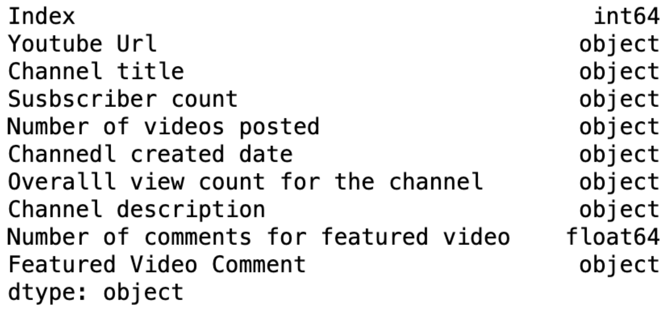


	Index	Instructor rating	Number of courses
count	3055.00	3055.00	2938.00
mean	1528.00	4.34	72.11
std	882.05	0.26	159.37
min	1.00	2.80	2.00
25%	764.50	4.20	7.00
50%	1528.00	4.40	19.00
75%	2291.50	4.50	46.00
max	3055.00	5.00	902.00

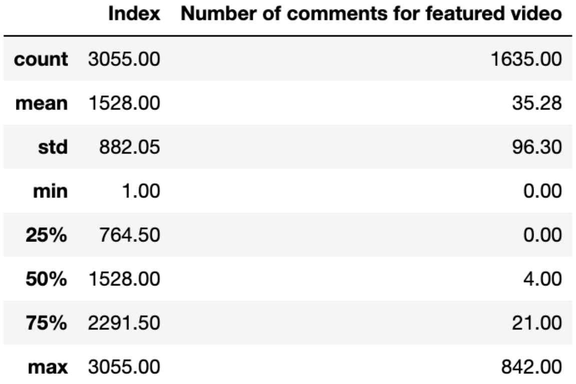
YouTube dataset obtained by web scraping from the above Udemy dataset



```
Number of rows and columns in data set: (3055, 10)
```

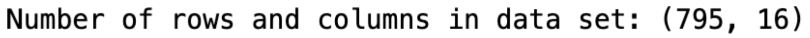


```
Index                                    int64
Youtube Url                             object
Channel title                           object
Susbscriber count                       object
Number of videos posted                 object
Channedl created date                   object
Overalll view count for the channel     object
Channel description                     object
Number of comments for featured video  float64
Featured Video Comment                  object
dtype: object
```

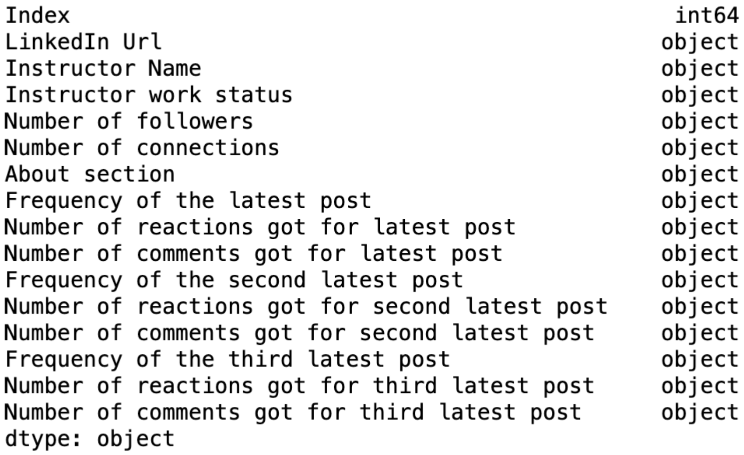


	Index	Number of comments for featured video
count	3055.00	1635.00
mean	1528.00	35.28
std	882.05	96.30
min	1.00	0.00
25%	764.50	0.00
50%	1528.00	4.00
75%	2291.50	21.00
max	3055.00	842.00

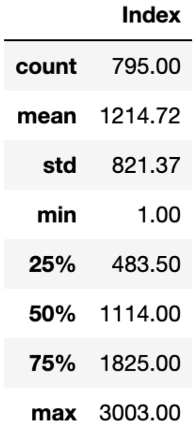
LinkedIn Individual Profiles dataset which was scraped from the cleaned Udemy dataset



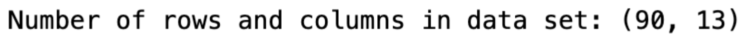
```
Number of rows and columns in data set: (795, 16)
```



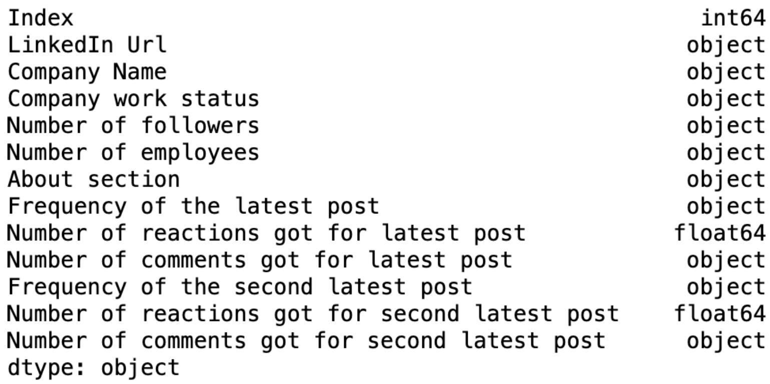
```
Index                                         int64
LinkedIn Url                                 object
Instructor Name                              object
Instructor work status                       object
Number of followers                          object
Number of connections                        object
About section                                object
Frequency of the latest post                 object
Number of reactions got for latest post      object
Number of comments got for latest post       object
Frequency of the second latest post          object
Number of reactions got for second latest post object
Number of comments got for second latest post  object
Frequency of the third latest post           object
Number of reactions got for third latest post  object
Number of comments got for third latest post   object
dtype: object
```



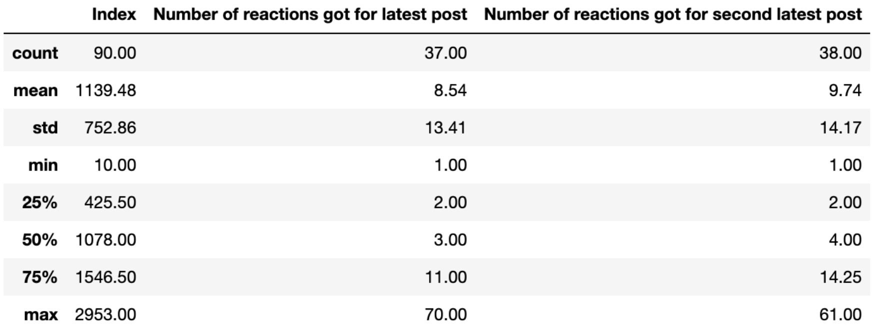
	Index
count	795.00
mean	1214.72
std	821.37
min	1.00
25%	483.50
50%	1114.00
75%	1825.00
max	3003.00



```
Number of rows and columns in data set: (90, 13)
```



```
Index                                               int64
LinkedIn Url                                       object
Company Name                                       object
Company work status                                object
Number of followers                                object
Number of employees                                object
About section                                      object
Frequency of the latest post                       object
Number of reactions got for latest post           float64
Number of comments got for latest post             object
Frequency of the second latest post                object
Number of reactions got for second latest post    float64
Number of comments got for second latest post      object
dtype: object
```



	Index	Number of reactions got for latest post	Number of reactions got for second latest post
count	90.00	37.00	38.00
mean	1139.48	8.54	9.74
std	752.86	13.41	14.17
min	10.00	1.00	1.00
25%	425.50	2.00	2.00
50%	1078.00	3.00	4.00
75%	1546.50	11.00	14.25
max	2953.00	70.00	61.00

Methodology

Step 1: Infrastructure Selection

- In our project, we employed Beautiful Soup and Selenium with ChromeDriver in Jupyter Notebook for dynamic data scraping. Beautiful Soup parsed HTML structures, while Selenium automated interactions with dynamic web pages. The extracted data was stored securely in AWS S3, ensuring centralized access, and facilitating integration into subsequent project stages.
- In our data cleaning phase, AWS Glue simplified the preparation and transformation of our scraped data, ensuring quality and consistency. The cleansed data was efficiently stored back in AWS S3.
- For performing sentiment analysis on youtube comments data, we used a textblob object to identify polarity of sentence and word cloud to show frequency of words in sentence. Latent Dirichlet Allocation(LDA) technique is used for topic modeling on the "About" section of LinkedIn profile to extract topics.

- We employed multiple linear regression using Python within the Jupyter Notebook environment. This statistical technique, facilitated by libraries like NumPy, Pandas, and Scikit-learn, allowed us to analyze relationships between multiple predictor variables and a response variable. The interactive and collaborative nature of Jupyter Notebook enhanced our model development, visualization, and interpretation processes, ensuring transparency and reproducibility in our multiple linear regression analysis.
- For analysis and visualization, we utilized AWS Athena and Tableau. AWS Athena enabled seamless querying of our data stored in AWS S3, while Tableau facilitated the creation of dynamic and insightful visualizations. These tools synergistically empowered us to derive actionable insights and present findings through compelling dashboards.

Step 2: Data collection

We possess a collection of 8000 Udemy URLs, primarily comprising links to business courses from Kaggle. Through web scraping these URLs, we can extract information about the instructors, including their social media profiles. Our primary focus was on platforms such as YouTube and LinkedIn. Extracted data, including text and links, was converted to a data frame, and stored as csv file. Automation enables efficient data collection across multiple pages. Legal and ethical considerations are paramount, requiring adherence to a website's terms of service.

Data extraction from Udemy:

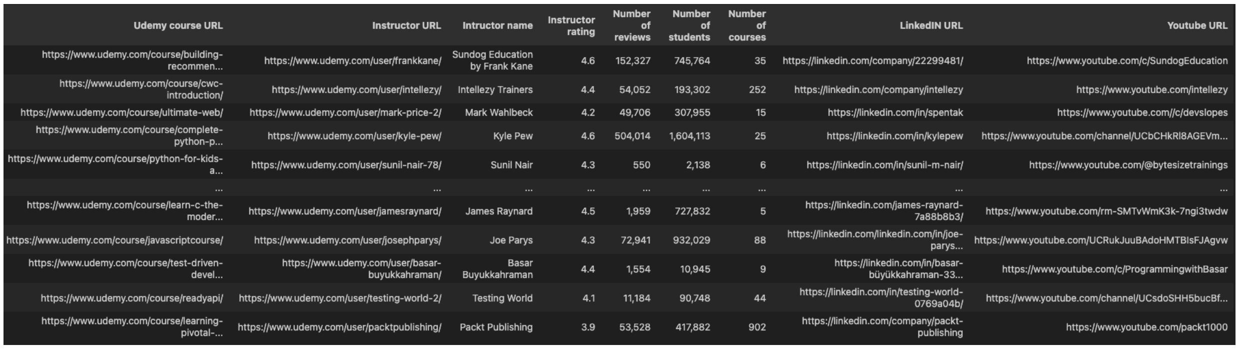
In our data extraction phase from Udemy, we leverage web scraping tools such as Beautiful Soup and Selenium, implemented in Python. The focus is on extracting comprehensive information about instructors, specifically targeting their social media profiles on platforms like YouTube and LinkedIn.

From a list of 8000 Udemy course URLs, we aim to retrieve data related to the instructors associated with the courses. Using web scraping, we systematically gather key information, including the Instructor URL, Name, Rating, Number of Reviews, Number of Students, and Number of Courses taught on Udemy.

To execute this, we initiate HTTP requests to the provided Udemy course URLs, performing initial requests and HTML parsing to identify instructor-related details. Python libraries, such as requests and Beautiful Soup, prove instrumental in extracting relevant data. We have included robust error-handling mechanisms, encapsulated in try-catch blocks, to address potential challenges like timeouts and inconsistencies in the web pages. In case of a timeout error, we provide a notification message, ensuring that the scraping process remains resilient.

Furthermore, to enhance data quality, a filter is applied to include entries featuring both YouTube and LinkedIn profile URLs. This filter contributes to refining the dataset and ensuring that the extracted information aligns with our specific criteria.

The part of the data frame created using the extracted data is provided below.



Udemy course URL	Instructor URL	Intructor name	Instructor rating	Number of reviews	Number of students	Number of courses	LinkedIN URL	Youtube URL
https://www.udemy.com/course/building-recommen...	https://www.udemy.com/user/frankkane/	Sundog Education by Frank Kane	4.6	152,327	745,764	35	https://linkedin.com/company/22299481/	https://www.youtube.com/c/SundogEducation
https://www.udemy.com/course/cwc-introduction/	https://www.udemy.com/user/intellezy/	Intellezy Trainers	4.4	54,052	193,302	252	https://linkedin.com/company/intellezy	https://www.youtube.com/intellezy
https://www.udemy.com/course/ultimate-web/	https://www.udemy.com/user/mark-price-2/	Mark Wahlbeck	4.2	49,706	307,955	15	https://linkedin.com/in/spentak	https://www.youtube.com//c/devslopes
https://www.udemy.com/course/complete-python-p...	https://www.udemy.com/user/kyle-pew/	Kyle Pew	4.6	504,014	1,604,113	25	https://linkedin.com/in/kylepew	https://www.youtube.com/channel/UCbCHkRl8AGEVm...
https://www.udemy.com/course/python-for-kids-a...	https://www.udemy.com/user/sunil-nair-78/	Sunil Nair	4.3	550	2,138	6	https://linkedin.com/in/sunil-m-nair/	https://www.youtube.com/@bytesizetrainings
...	...	...	...	...	...	...	...	...
https://www.udemy.com/course/learn-c-the-moder...	https://www.udemy.com/user/jamesraynard/	James Raynard	4.5	1,959	727,832	5	https://linkedin.com/james-raynard-7a88b8b3/	https://www.youtube.com/rm-SMTvWmK3k-7ngi3twdw
https://www.udemy.com/course/javascriptcourse/	https://www.udemy.com/user/josephparys/	Joe Parys	4.3	72,941	932,029	88	https://linkedin.com/linkedin.com/in/joe-parys...	https://www.youtube.com/UCRukJuuBAdoHMTBIsFJAgvw
https://www.udemy.com/course/test-driven-devel...	https://www.udemy.com/user/basar-buyukkahraman/	Basar Buyukkahraman	4.4	1,554	10,945	9	https://linkedin.com/in/basar-büyükkahraman-33...	https://www.youtube.com/c/ProgrammingwithBasar
https://www.udemy.com/course/readyapi/	https://www.udemy.com/user/testing-world-2/	Testing World	4.1	11,184	90,748	44	https://linkedin.com/in/testing-world-0769a04b/	https://www.youtube.com/channel/UCsdoSHH5bucBf...
https://www.udemy.com/course/learning-pivotal-...	https://www.udemy.com/user/packtpublishing/	Packt Publishing	3.9	53,528	417,882	902	https://linkedin.com/company/packt-publishing	https://www.youtube.com/packt1000

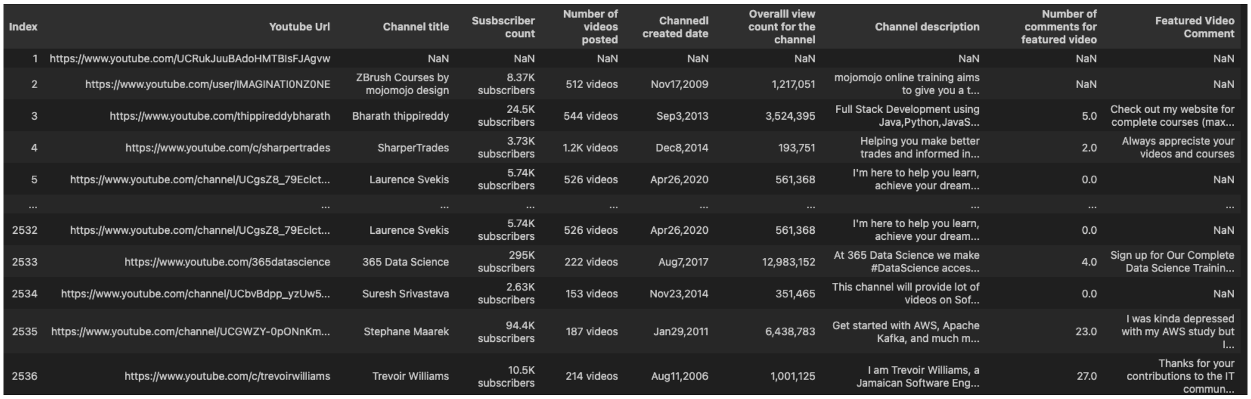
Data extraction from YouTube

We utilize a loop structure to iteratively process a list of YouTube channel URLs (which we have extracted in Udemy scraping), employing Selenium as the web automation framework. This systematic approach allows for the extraction of critical channel metrics, such as title, subscriber count, video count, creation date, and view count. Selenium initializes the WebDriver with Chrome, facilitating automated interactions with the YouTube web pages. The incorporation of implicit waiting enhances adaptability to varying loading times.

We have also incorporated error-handling mechanisms, efficiently managing exceptions encountered during web automation. In case of an unexpected error, the affected data is gracefully marked as "N/A," ensuring the continuous execution of the scraping process. Extending its functionality, we have also extracted top 20 comments from featured videos. It employs simulated scrolling actions using ActionChains(driver) and interacts with the YouTube page to load comments dynamically. We have captured both the comments and their counts, maintaining an organized structure for further analysis.

To manage potential timeouts, we incorporate an implicit timeout mechanism, setting a waiting period of 5 to 60 seconds for page loading. This proactive approach ensures that the script gracefully handles timeouts, providing users with timely notifications and allowing for additional error-handling strategies.

The part of data frame created using the extracted data is provided below.



Index	Youtube Url	Channel title	Susbscriber count	Number of videos posted	Channedl created date	Overalll view count for the channel	Channel description	Number of comments for featured video	Featured Video Comment
1	https://www.youtube.com/UCRukJuuBAdoHMTBIsFJAgvw	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	https://www.youtube.com/user/IMAGINATI0NZ0NE	ZBrush Courses by mojomojo design	8.37K subscribers	512 videos	Nov17,2009	1,217,051	mojomojo online training aims to give you a t...	NaN	NaN
3	https://www.youtube.com/thippireddybharath	Bharath thippireddy	24.5K subscribers	544 videos	Sep3,2013	3,524,395	Full Stack Development using Java,Python,JavaS...	5.0	Check out my website for complete courses (max...
4	https://www.youtube.com/c/sharpertrades	SharperTrades	3.73K subscribers	1.2K videos	Dec8,2014	193,751	Helping you make better trades and informed in...	2.0	Always appreciste your videos and courses
5	https://www.youtube.com/channel/UCgsZ8_79Eclct...	Laurence Svekis	5.74K subscribers	526 videos	Apr26,2020	561,368	I'm here to help you learn, achieve your dream...	0.0	NaN
...	...	...	...	...	...	...	...	...	...
2532	https://www.youtube.com/channel/UCgsZ8_79Eclct...	Laurence Svekis	5.74K subscribers	526 videos	Apr26,2020	561,368	I'm here to help you learn, achieve your dream...	0.0	NaN
2533	https://www.youtube.com/365datascience	365 Data Science	295K subscribers	222 videos	Aug7,2017	12,983,152	At 365 Data Science we make #DataScience acces...	4.0	Sign up for Our Complete Data Science Trainin...
2534	https://www.youtube.com/channel/UCbvBdpp_yzUw5...	Suresh Srivastava	2.63K subscribers	153 videos	Nov23,2014	351,465	This channel will provide lot of videos on Sof...	0.0	NaN
2535	https://www.youtube.com/channel/UCGWZY-0pONnKm...	Stephane Maarek	94.4K subscribers	187 videos	Jan29,2011	6,438,783	Get started with AWS, Apache Kafka, and much m...	23.0	I was kinda depressed with my AWS study but I...
2536	https://www.youtube.com/c/trevoirwilliams	Trevoir Williams	10.5K subscribers	214 videos	Aug11,2006	1,001,125	I am Trevoir Williams, a Jamaican Software Eng...	27.0	Thanks for your contributions to the IT commun...

Data extraction from LinkedIn

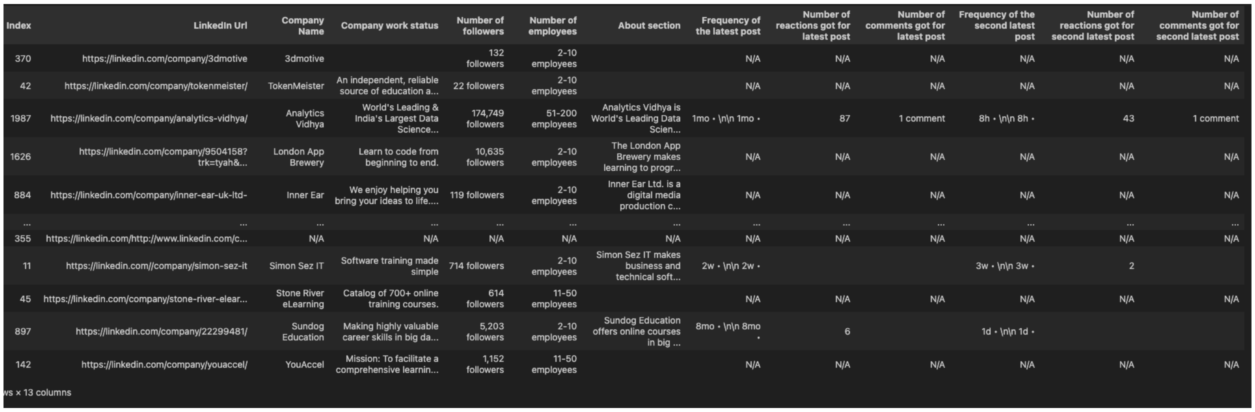
The LinkedIn scraping script adeptly navigates through two types of pages: company pages and personal pages. Starting with a list of LinkedIn profile URLs obtained from Udemy scraping, the script intelligently excludes uninformative or empty URLs using an exclusion set.

For personal pages, Selenium WebDriver efficiently opens each LinkedIn profile URL, gracefully handling potential errors and marking specific data points as "N/A" when needed. Extracted data includes the instructor's name, current workplace, number of connections, followers, and any available "About" sections. Additionally, we delve into recent LinkedIn posts, capturing details on posting frequency, reactions, and comments for the top three posts.

We employ try-except blocks to handle exceptions gracefully, ensuring smooth data extraction even in the face of timeouts or unexpected errors. Informative messages like "Timeout occurred" or "Exception occurred" are printed to notify users about issues. Implicit waiting, explicit time delays, and exception handling collectively contribute to the script's effective management of timeouts, ensuring a seamless data extraction process.

For LinkedIn company pages, the script iterates through a list of company profile URLs, leveraging Selenium WebDriver for data extraction. Similar to personal pages, it interacts with web pages, retrieves data, and adeptly handles potential issues, maintaining consistency with the approach used for LinkedIn personal page extraction.

The part of data frame created using the extracted data is provided below.



Index	LinkedIn Url	Company Name	Company work status	Number of followers	Number of employees	About section	Frequency of the latest post	Number of reactions got for latest post	Number of comments got for latest post	Frequency of the second latest post	Number of reactions got for second latest post	Number of comments got for second latest post
370	https://linkedin.com/company/3dmotive	3dmotive		132 followers	2-10 employees		N/A	N/A	N/A	N/A	N/A	N/A
42	https://linkedin.com/company/tokenmeister/	TokenMeister	An independent, reliable source of education a...	22 followers	2-10 employees		N/A	N/A	N/A	N/A	N/A	N/A
1987	https://linkedin.com/company/analytics-vidhya/	Analytics Vidhya	World's Leading & India's Largest Data Science...	174,749 followers	51-200 employees	Analytics Vidhya is World's Leading Data Scien...	1mo • \n\n 1mo •	87	1 comment	8h • \n\n 8h •	43	1 comment
1626	https://linkedin.com/company/9504158?trk=tyah&...	London App Brewery	Learn to code from beginning to end.	10,635 followers	2-10 employees	The London App Brewery makes learning to progr...	N/A	N/A	N/A	N/A	N/A	N/A
884	https://linkedin.com/company/inner-ear-uk-ltd-	Inner Ear	We enjoy helping you bring your ideas to life....	119 followers	2-10 employees	Inner Ear Ltd. is a digital media production c...	N/A	N/A	N/A	N/A	N/A	N/A
...	...	...	...	...	...	...	...	...	...	...	...	...
355	https://linkedin.com/http://www.linkedin.com/c...	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
11	https://linkedin.com//company/simon-sez-it	Simon Sez IT	Software training made simple	714 followers	2-10 employees	Simon Sez IT makes business and technical soft...	2w • \n\n 2w •			3w • \n\n 3w •	2	
45	https://linkedin.com/company/stone-river-elear...	Stone River eLearning	Catalog of 700+ online training courses.	614 followers	11-50 employees		N/A	N/A	N/A	N/A	N/A	N/A
897	https://linkedin.com/company/22299481/	Sundog Education	Making highly valuable career skills in big da...	5,203 followers	2-10 employees	Sundog Education offers online courses in big ...	8mo • \n\n 8mo •	6		1d • \n\n 1d •		
142	https://linkedin.com/company/youaccel/	YouAccel	Mission: To facilitate a comprehensive learnin...	1,152 followers	11-50 employees		N/A	N/A	N/A	N/A	N/A	N/A

ws × 13 columns

Step 3: Data Cleaning

In the LinkedIn, Udemy, and YouTube datasets, we encountered challenges such as missing values, inconsistent formatting, and special characters. AWS Glue DataBrew proved instrumental in simplifying the data preparation process, serving as a transformative tool.

Data Cleaning Steps using AWS Glue DataBrew

A. Creating a Dataset.

In DataBrew, a dataset is an organized library, neatly sorting data into rows and columns. Creating datasets—Udemy, YouTube, and LinkedIn—began by uploading CSV files from S3, laying the foundation for seamless transformations.

B. Data Profile Job.

Data profiling swiftly generates a detailed summary report for a dataset. Running a data profile job reveals key details about the dataset's shape, content, structure, and relationships, providing swift insights for analysis.

After creating a dataset, we can run a data profile job to understand the structure of our data. The output of this job will be saved to Amazon S3. The data profile job is initiated by providing a job profile name, specifying the S3 path, and configuring additional settings such as dataset-level configurations or column-level configurations. Additionally, we set up the required role name and permissions.

In the Udemy dataset, we started with 3055 pieces of information organized in 10 categories (columns). There were 117 empty spots in our data. To understand it better, we looked at correlation, box plots for outliers, a summary of each column, and the types of data in them.
This process was done for all three datasets to thoroughly explore our data.

C. Creating a Project.

A project is the main focus of the work when analyzing and transforming data.
Beginning by establishing a Udemy project, connecting it to the dataset generated in the previous step. This involved assigning a project name, selecting the dataset, specifying the desired number of rows for our sample data, and configuring the necessary permissions and IAM roles.
Three projects were created, one for each dataset. After making the project, we go into the workspace to clean up the dataset. In the workspace, we create recipes to tidy up the data.

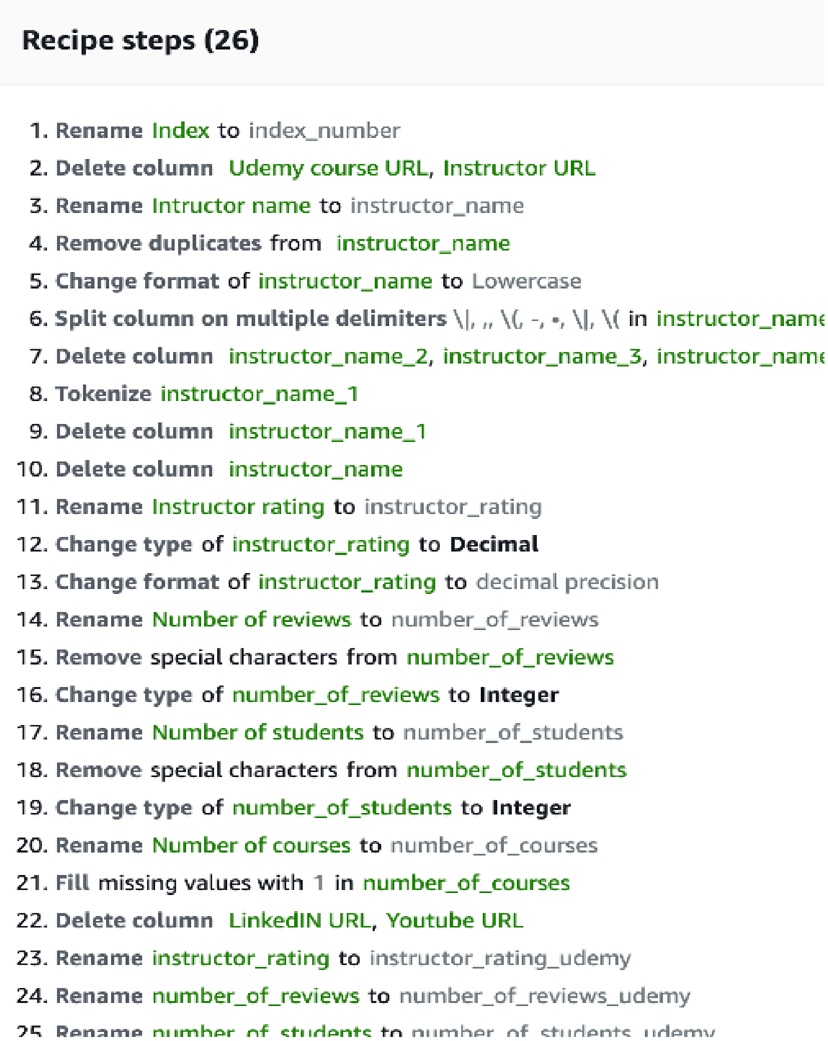
D. Creating Recipes.

A DataBrew recipe provided step-by-step instructions for transformations. DataBrew also retains the recipe details, not the data itself. We can easily download and reuse recipes in different projects, creating variations when needed. Recipes were created for each of the dataset.

Udemy Data Cleaning.

The data cleaning process for the Udemy Courses dataset involved several key steps to enhance its quality and structure. The index was renamed to "index_number," and irrelevant columns such as "Udemy Course URL," "Instructor URL," "LinkedIn URL," and "Youtube URL" were deleted for streamlining.
The "Instructor Name" column underwent a comprehensive refinement, including renaming, removal of duplicates, and splitting based on delimiters like |,.,(to extract relevant information. After converting the values to lowercase, the names were joined using an underscore and stored in the destination column "instructor_name." Temporary deletion of the intermediate "instructor_name" column facilitated regression.
For "Instructor Rating," the column was renamed, and its data type was changed to decimal with a precision of 1. The "Number of Reviews" and "Number of Students" columns underwent renaming, removal of special characters, and changes in data type to integers. The "Number of

Courses" column was renamed, and missing values were filled with 1 to address data scraping issues.



Recipe steps (26)

1. Rename Index to index_number
2. Delete column Udemy course URL, Instructor URL
3. Rename Intructor name to instructor_name
4. Remove duplicates from instructor_name
5. Change format of instructor_name to Lowercase
6. Split column on multiple delimiters \|, ,, \(, -, •, \|, \(in instructor_name
7. Delete column instructor_name_2, instructor_name_3, instructor_name
8. Tokenize instructor_name_1
9. Delete column instructor_name_1
10. Delete column instructor_name
11. Rename Instructor rating to instructor_rating
12. Change type of instructor_rating to **Decimal**
13. Change format of instructor_rating to decimal precision
14. Rename Number of reviews to number_of_reviews
15. Remove special characters from number_of_reviews
16. Change type of number_of_reviews to **Integer**
17. Rename Number of students to number_of_students
18. Remove special characters from number_of_students
19. Change type of number_of_students to **Integer**
20. Rename Number of courses to number_of_courses
21. Fill missing values with 1 in number_of_courses
22. Delete column LinkedIN URL, Youtube URL
23. Rename instructor_rating to instructor_rating_udemy
24. Rename number_of_reviews to number_of_reviews_udemy
25. Rename number_of_students to number_of_students_udemy

YouTube Data Cleaning.

The data cleaning process for the YouTube dataset involved meticulous steps to enhance its quality and structure. The index was renamed to "index_number," and the YouTube URL column was deleted. The "Channel Title" column underwent a series of transformations, including renaming, removal of duplicates and empty rows, conversion to lowercase, tokenization, and eventual deletion.

For the "Subscriber Count" column, custom values were removed, suffixes denoting thousands (K) and millions (M) were addressed, and a new column, "total_subscriber_count," was created by multiplying two relevant columns. The "Number of Videos posted" column underwent a similar process, involving renaming, removal of suffixes, conversion of data types, and the creation of a new column, "total_videos_posted."

The "Channel Created Date" column was cleaned by splitting and categorically mapping the month, resulting in a standardized format of YYYY-MM-DD. The "Overall view count for the channel" was renamed, and commas were removed, with missing values handled appropriately.

The "Channel Description" column was renamed to "channel_description." The "Number of comments on featured video" column was renamed, missing values were filled with zeros, and its data type was changed to integer. The "Featured Video comment" column was split into 20 columns using a specific delimiter.

These comprehensive steps ensure a refined and structured YouTube dataset, laying the foundation for meaningful analysis and insights.

LinkedIn Data Cleaning.

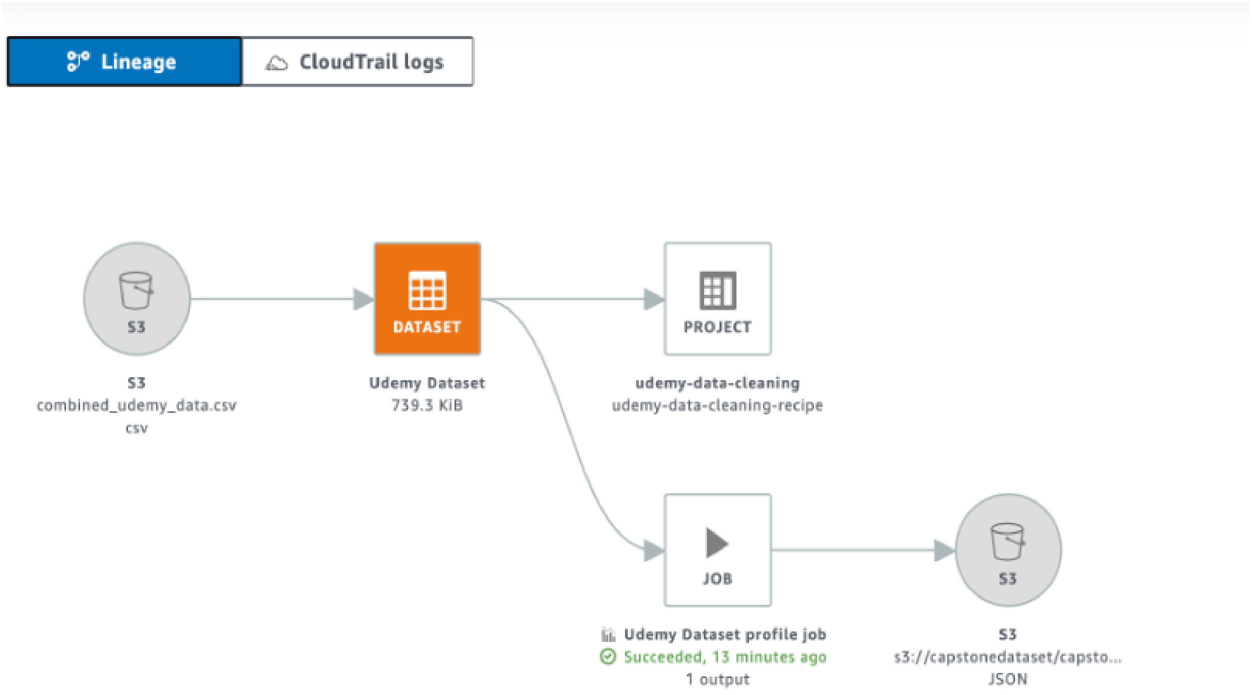
The data cleaning process for the LinkedIn Instructor dataset involved several key steps. First, adjustments were made to the index and LinkedIn URL columns. The instructor's name underwent a thorough refinement process, including the removal of null values, duplicates, and special characters, as well as a transformation to lowercase and tokenization. The instructor's work status column was permanently deleted, while the number of followers and connections columns were standardized by removing custom values, special characters, and converting data types.

Additionally, a new column, "number_of_connections_mapped," was created to categorize the connection counts. The "About" section underwent renaming and temporary deletion for regression purposes. The counts of reactions and comments for the latest three posts were processed similarly, with null values replaced, special characters removed, and data types adjusted.

Furthermore, the creation of new columns, such as "average_reaction_count" and "average_comment_count," streamlined the dataset. Lastly, the frequency of the latest three posts was addressed, with null values removed and time granularities standardized to days for consistency. The average frequency across these columns was then computed and added to the dataset. Overall, these meticulous steps ensured a clean and standardized dataset for further analysis of LinkedIn Instructors.

E. Data Lineage.

Data Lineage functions as a visual map within DataBrew, systematically tracking the trajectory of your data. This view, referred to as data lineage, provides insights into the origin of your data, its interactions with various entities, temporal transformations, and its ultimate storage location. In the Udemy dataset journey, we started by bringing in the file from S3, made a dataset and a project. We then used data cleaning steps, checked the dataset profile, and saved the results to S3 after running the profile job.

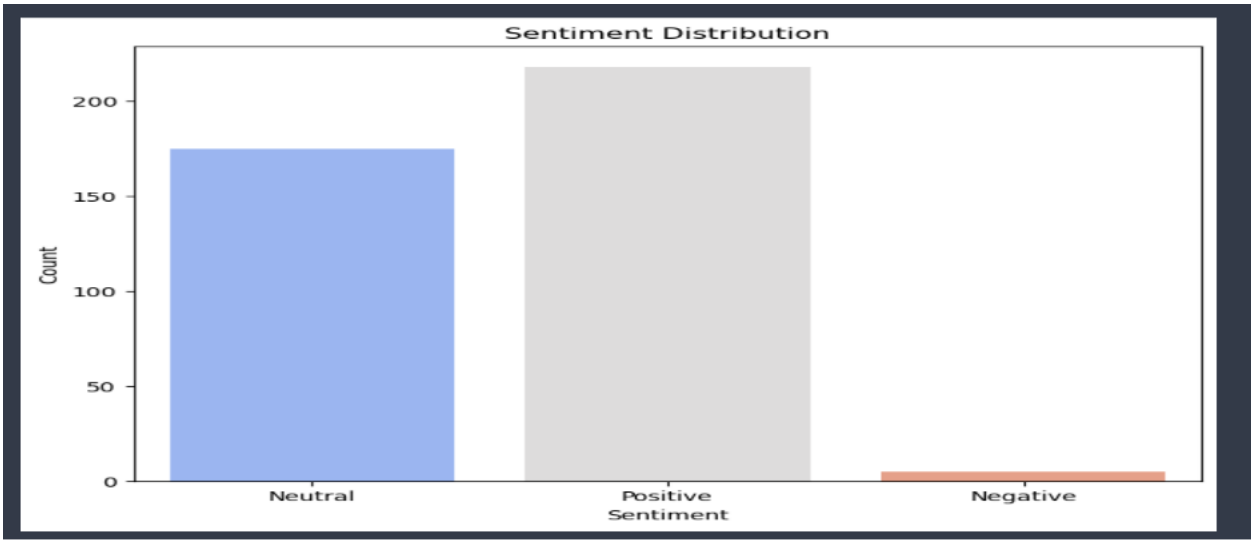


Step 4: Sentiment Analysis and Text Mining

Sentiment analysis is the process of determining the emotional tone behind a piece of text, such as a comment, tweet, or review. It involves classifying the sentiment of the text into categories like positive, negative, or neutral. YouTube, the world's largest video-sharing platform, is a treasure trove of comments on an array of topics. Analyzing the sentiments expressed in these comments can provide valuable insights into audience reactions, content creators' performance, and audience engagement.

Sentiment Analysis using TextBlob.

The TextBlob library was employed for sentiment analysis due to its simplicity and effectiveness in determining the polarity of textual data. The sentiment analysis function classified comments into three categories: Positive, Negative, and Neutral based on the polarity scores provided by TextBlob.



The sentiment analysis using TextBlob revealed the distribution of sentiments in the YouTube comments dataset. The comments were categorized as follows:

Positive: Expressing a positive sentiment.

Negative: Expressing a negative sentiment.

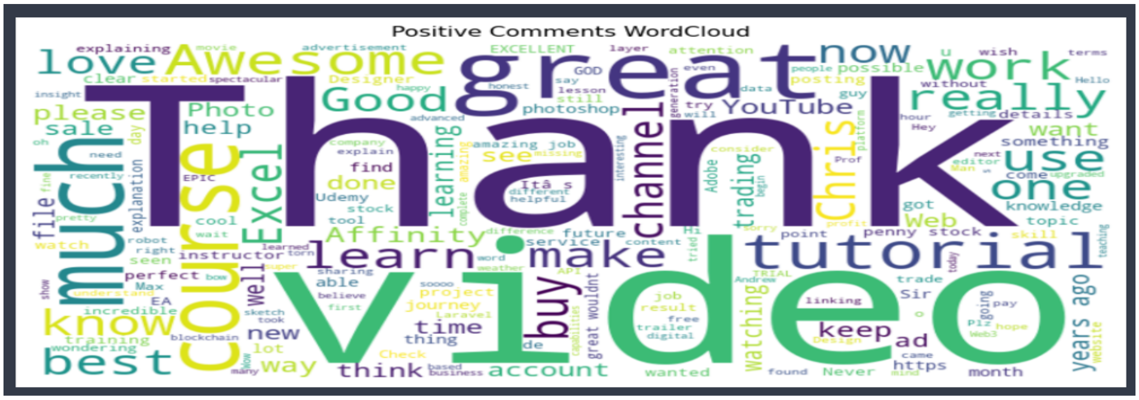
Neutral: Neither positive nor negative in sentiment.

Sentiment Analysis using WordCloud Generation

WordClouds were generated to provide a visual representation of the most frequently occurring words in both positive and negative comments. WordClouds are useful in identifying prominent terms that encapsulate the sentiments expressed in the comments.

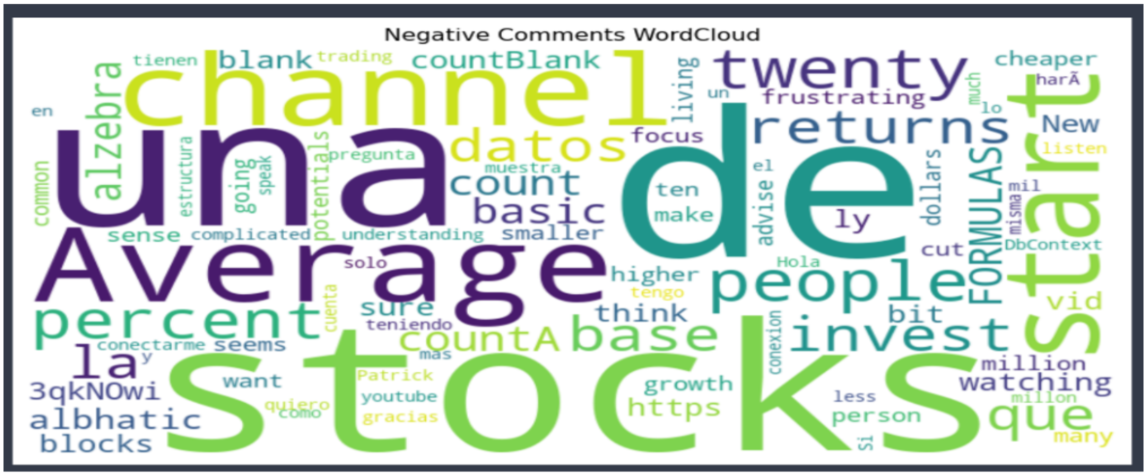
Positive Comments WordCloud.

The WordCloud for positive comments visually highlights the words that frequently appear in comments with positive sentiments. Larger fonts indicate higher frequency.



Negative Comments WordCloud.

The WordCloud for negative comments illustrates the most common terms found in comments with negative sentiments. Larger fonts represent higher frequency.



Topic modeling on LinkedIn "About" information

One powerful technique to enhance the effectiveness of our LinkedIn "About" section is topic modeling, a method in natural language processing that uncovers hidden thematic structures within a text. By employing algorithms to analyze content and identify recurring themes or topics, topic modeling not only organizes information but also boosts the discoverability of key skills, experiences, and interests.

Below are the steps used to implement topic modeling.

1. Imported all the required libraries and packages.
2. Imported cleaned LinkedIn CSV data file containing attributes such as follower count, number of connections, and the "about_section_linkedin" information.

3. Removal of stop words and stemming**:**
   Stop words removal involves eliminating common, non-substantive words from text data to enhance computational efficiency in topic modeling.Stemming reduces words to their root form, aiding in grouping similar terms but potentially sacrificing semantic precision.
4. Train the Latent Dirichlet Allocation (LDA) model:
   LDA is a generative probabilistic model widely used in topic modeling. It assumes that documents are mixtures of topics, and topics are mixtures of words.LDA helps unveil underlying thematic structures and extract meaningful insights from large datasets.
5. Selection of optimal number of topics using coherence scores:
   The selection of the number of topics and words for topic modeling depends on the nature of our data, our research objectives, and the interpretability of the results.
   Coherence score is used to evaluate the quality of topics for different numbers.

Example of topic modeling

Text: “I am passionate about studying and teaching”

Result:

```
[(0, '0.538*"studi" + 0.306*"teach" + 0.078*"passion" + 0.078*"teachingi"'), (1,
'0.566*"teachingi" + 0.145*"passion" + 0.145*"teach" + 0.145*"studi"'), (2,
'0.697*"passion" + 0.101*"studi" + 0.101*"teach" + 0.101*"teachingi"')]
```

Topic 0: **"Passion for Studying and Teaching"**

Keyword: "studi," "teach," "passion," "teachingi"

Interpretation: This topic seems to revolve around a combination of studying, teaching, and expressing passion for these activities.

Topic 1: **"Teaching and Passion"**

Keyword: "teachingi,""passion,""teach," "studi"

Interpretation: This topic places a higher emphasis on the term "teachingi" compared to the other keywords, suggesting a focus on a particular aspect of teaching or instruction, along with a passion for it.

Topic 2: **"Passion in Education"**

Keyword: "passion," "studi," "teach," "teachingi"

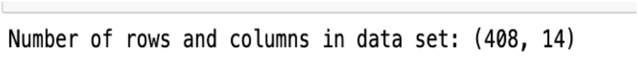
Interpretation: This topic is characterized by a strong emphasis on the term "passion," suggesting a topic related to expressing enthusiasm or strong feelings, possibly in the context of studying and teaching.

Step 5: Data Analysis

The objective of this analysis is to investigate the factors influencing the number of students on Udemy, utilizing a dataset that combines information from Udemy, YouTube, and LinkedIn instructor profiles. The analysis employs multiple linear regression to explore the relationships between various predictor variables and the target variable, **'number_of_students_udemy**'. Initially, we import the cleansed dataset using the pandas library and then proceed with data exploration.

Data Overview

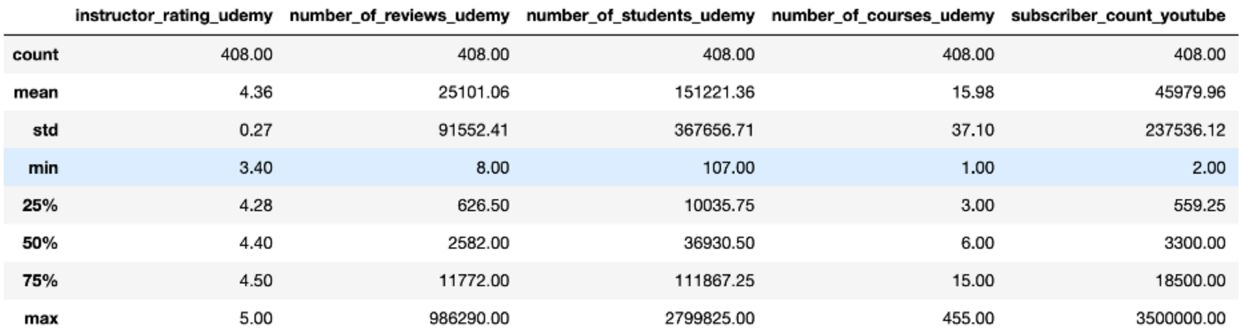
The dataset consists of three main components: Udemy, YouTube, and LinkedIn instructor data. The datasets were merged based on a common key (**'index_number'**) and sorted in ascending order. The dimensions of the combined dataset are:



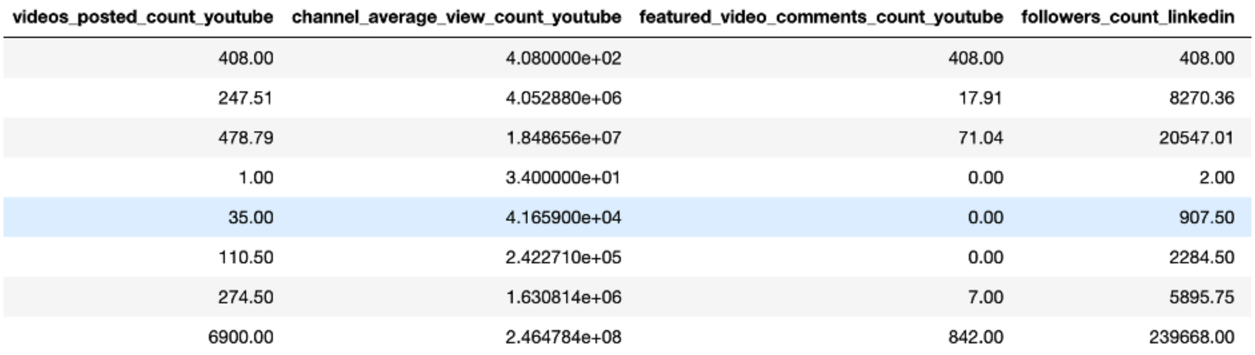
```
Number of rows and columns in data set: (408, 14)
```

Descriptive Statistics

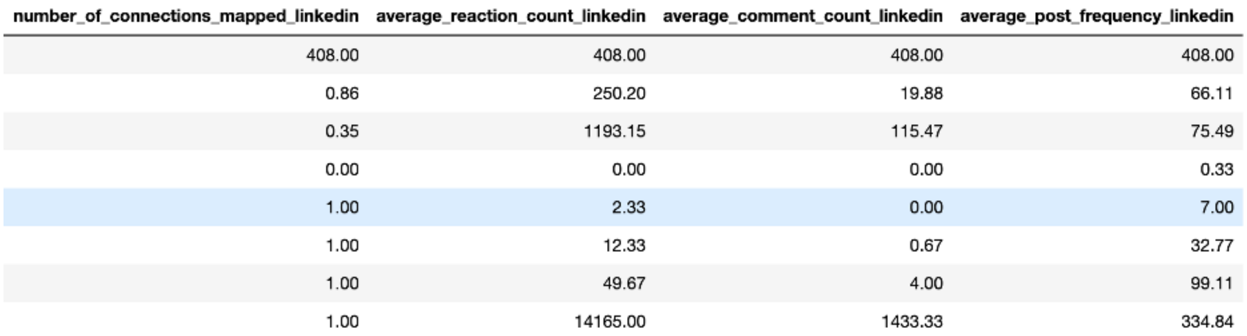
In this analysis, descriptive statistics were calculated to understand the basic properties of the transformed dataset.



	instructor_rating_udemy	number_of_reviews_udemy	number_of_students_udemy	number_of_courses_udemy	subscriber_count_youtube
count	408.00	408.00	408.00	408.00	408.00
mean	4.36	25101.06	151221.36	15.98	45979.96
std	0.27	91552.41	367656.71	37.10	237536.12
min	3.40	8.00	107.00	1.00	2.00
25%	4.28	626.50	10035.75	3.00	559.25
50%	4.40	2582.00	36930.50	6.00	3300.00
75%	4.50	11772.00	111867.25	15.00	18500.00
max	5.00	986290.00	2799825.00	455.00	3500000.00



videos_posted_count_youtube	channel_average_view_count_youtube	featured_video_comments_count_youtube	followers_count_linkedin
408.00	4.080000e+02	408.00	408.00
247.51	4.052880e+06	17.91	8270.36
478.79	1.848656e+07	71.04	20547.01
1.00	3.400000e+01	0.00	2.00
35.00	4.165900e+04	0.00	907.50
110.50	2.422710e+05	0.00	2284.50
274.50	1.630814e+06	7.00	5895.75
6900.00	2.464784e+08	842.00	239668.00



number_of_connections_mapped_linkedin	average_reaction_count_linkedin	average_comment_count_linkedin	average_post_frequency_linkedin
408.00	408.00	408.00	408.00
0.86	250.20	19.88	66.11
0.35	1193.15	115.47	75.49
0.00	0.00	0.00	0.33
1.00	2.33	0.00	7.00
1.00	12.33	0.67	32.77
1.00	49.67	4.00	99.11
1.00	14165.00	1433.33	334.84

Predictor Variables and Outcome Variable

The following predictor variables were identified for the regression model:

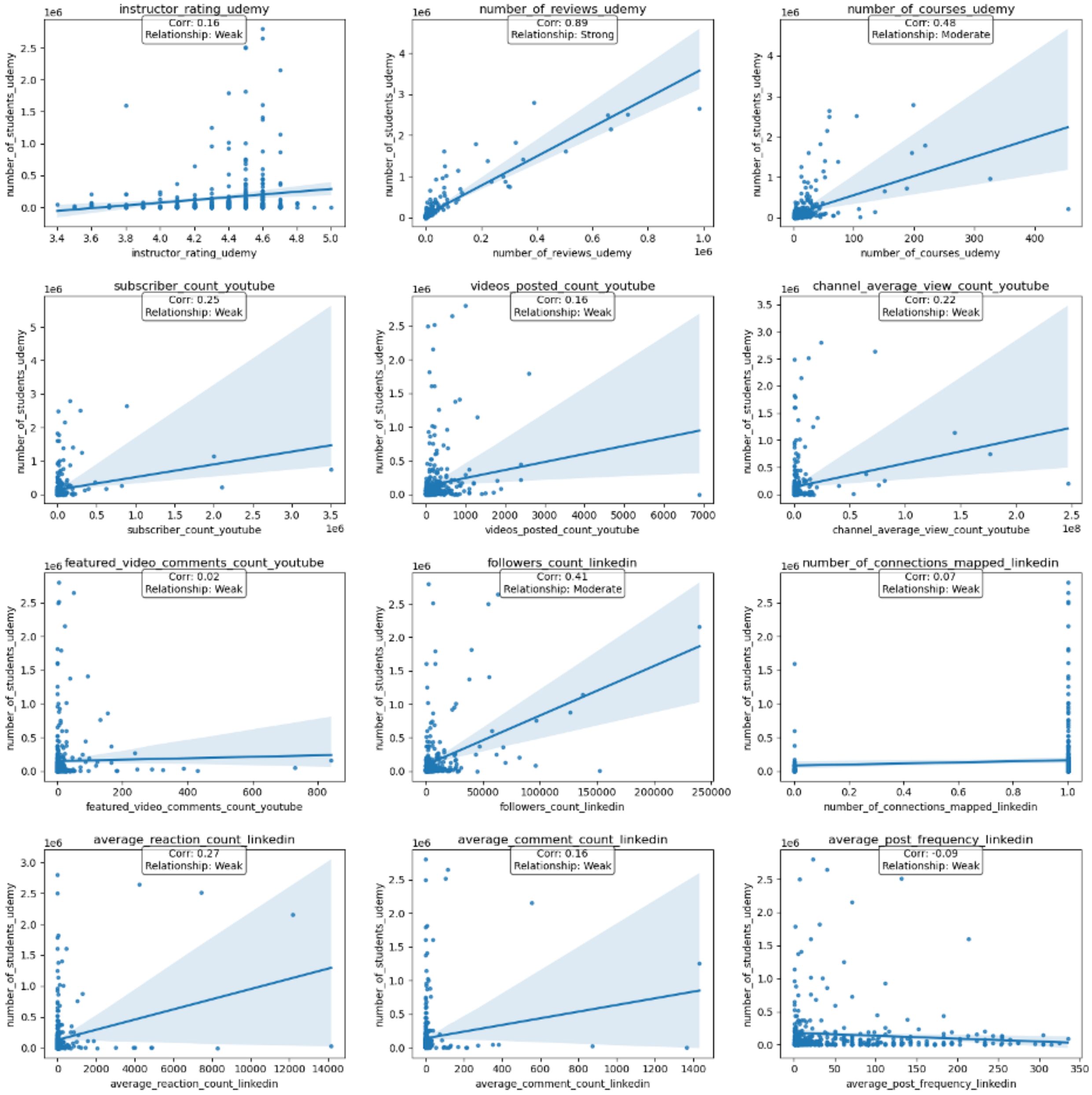
- Instructor Rating on Udemy
- Number of Reviews on Udemy
- Number of Courses on Udemy
- Subscriber Count on YouTube
- Videos Posted Count on YouTube
- Channel Average View Count on YouTube
- Featured Video Comments Count on YouTube
- Followers Count on LinkedIn
- Number of Connections Mapped on LinkedIn
- Average Reaction Count on LinkedIn
- Average Comment Count on LinkedIn
- Average Post Frequency on LinkedIn

The outcome variable is 'number_of_students_udemy.'

Assumption Checks

Linearity.

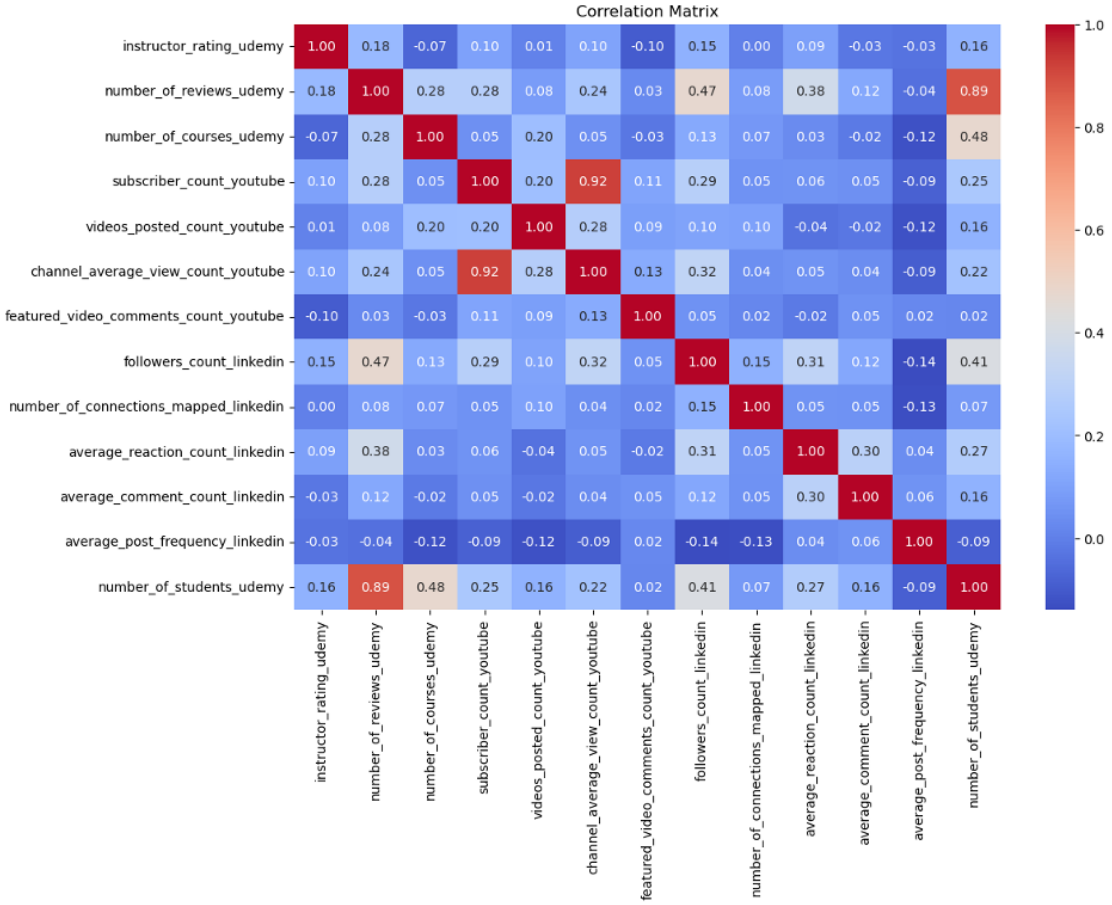
To assess linearity, scatter plots were created for each predictor against the outcome variable. The correlation matrix was also calculated to categorize relationships between variables as 'Weak,' 'Moderate,' or 'Strong.'



We observe that almost all the correlations are weak, except for the correlation between number of students and number of reviews, which exhibits a strong relationship and between number of students and followers count on LinkedIn which exhibits a moderate relationship.

Multicollinearity.

Multicollinearity was evaluated through a heatmap of the correlation matrix. High correlations among predictors could impact the stability of regression coefficients.



Correlation Matrix

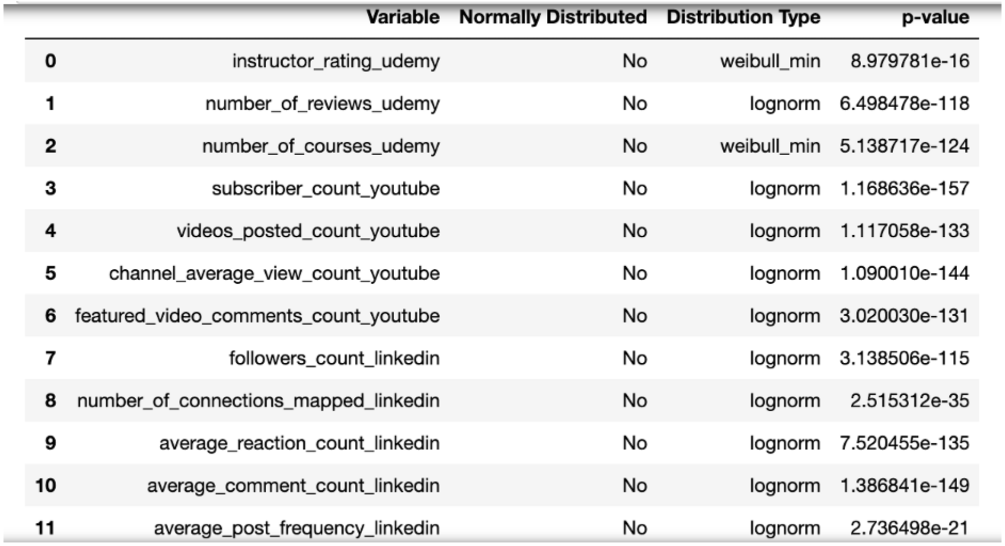
	instructor_rating_udemy	number_of_reviews_udemy	number_of_courses_udemy	subscriber_count_youtube	videos_posted_count_youtube	channel_average_view_count_youtube	featured_video_comments_count_youtube	followers_count_linkedin	number_of_connections_mapped_linkedin	average_reaction_count_linkedin	average_comment_count_linkedin	average_post_frequency_linkedin	number_of_students_udemy
instructor_rating_udemy	1.00	0.18	-0.07	0.10	0.01	0.10	-0.10	0.15	0.00	0.09	-0.03	-0.03	0.16
number_of_reviews_udemy	0.18	1.00	0.28	0.28	0.08	0.24	0.03	0.47	0.08	0.38	0.12	-0.04	0.89
number_of_courses_udemy	-0.07	0.28	1.00	0.05	0.20	0.05	-0.03	0.13	0.07	0.03	-0.02	-0.12	0.48
subscriber_count_youtube	0.10	0.28	0.05	1.00	0.20	0.92	0.11	0.29	0.05	0.06	0.05	-0.09	0.25
videos_posted_count_youtube	0.01	0.08	0.20	0.20	1.00	0.28	0.09	0.10	0.10	-0.04	-0.02	-0.12	0.16
channel_average_view_count_youtube	0.10	0.24	0.05	0.92	0.28	1.00	0.13	0.32	0.04	0.05	0.04	-0.09	0.22
featured_video_comments_count_youtube	-0.10	0.03	-0.03	0.11	0.09	0.13	1.00	0.05	0.02	-0.02	0.05	0.02	0.02
followers_count_linkedin	0.15	0.47	0.13	0.29	0.10	0.32	0.05	1.00	0.15	0.31	0.12	-0.14	0.41
number_of_connections_mapped_linkedin	0.00	0.08	0.07	0.05	0.10	0.04	0.02	0.15	1.00	0.05	0.05	-0.13	0.07
average_reaction_count_linkedin	0.09	0.38	0.03	0.06	-0.04	0.05	-0.02	0.31	0.05	1.00	0.30	0.04	0.27
average_comment_count_linkedin	-0.03	0.12	-0.02	0.05	-0.02	0.04	0.05	0.12	0.05	0.30	1.00	0.06	0.16
average_post_frequency_linkedin	-0.03	-0.04	-0.12	-0.09	-0.12	-0.09	0.02	-0.14	-0.13	0.04	0.06	1.00	-0.09
number_of_students_udemy	0.16	0.89	0.48	0.25	0.16	0.22	0.02	0.41	0.07	0.27	0.16	-0.09	1.00

Normality.

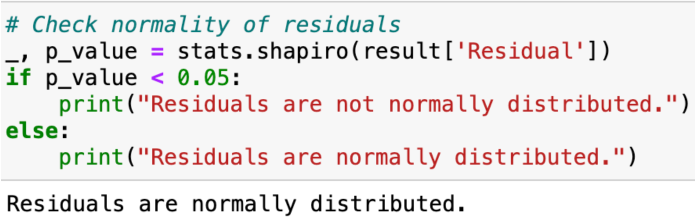
The normality of residuals was assessed using a Shapiro-Wilk test. Normality is a key assumption for the validity of regression results.

The null hypothesis of this test is that the data follows a normal distribution. If the p-value from the test is greater than or equal to 0.05, the variable is normally distributed; otherwise, it is labeled as non-normally distributed.

We can observe that the variables exhibit deviations from normality.



	Variable	Normally Distributed	Distribution Type	p-value
0	instructor_rating_udemy	No	weibull_min	8.979781e-16
1	number_of_reviews_udemy	No	lognorm	6.498478e-118
2	number_of_courses_udemy	No	weibull_min	5.138717e-124
3	subscriber_count_youtube	No	lognorm	1.168636e-157
4	videos_posted_count_youtube	No	lognorm	1.117058e-133
5	channel_average_view_count_youtube	No	lognorm	1.090010e-144
6	featured_video_comments_count_youtube	No	lognorm	3.020030e-131
7	followers_count_linkedin	No	lognorm	3.138506e-115
8	number_of_connections_mapped_linkedin	No	lognorm	2.515312e-35
9	average_reaction_count_linkedin	No	lognorm	7.520455e-135
10	average_comment_count_linkedin	No	lognorm	1.386841e-149
11	average_post_frequency_linkedin	No	lognorm	2.736498e-21



```python
# Check normality of residuals
_, p_value = stats.shapiro(result['Residual'])
if p_value < 0.05:
    print("Residuals are not normally distributed.")
else:
    print("Residuals are normally distributed.")
```

```
Residuals are normally distributed.
```

Note: We acknowledge that linear regression relies on several assumptions for its validity, aiming for a linear relationship between predictors and the response variable, independent and homoscedastic residuals, normality of residuals, absence of perfect multicollinearity, no autocorrelation, and additivity. While these assumptions are ideal for optimal model performance, it is important to recognize that real-world datasets may deviate from these conditions.

In practice, it is not always feasible for all assumptions to hold true. Variations in data distribution, outliers, and complex relationships between variables can contribute to deviations.

Data Splitting

To assess the model's generalizability, the dataset was divided into training and validation sets. The split utilized 80% of the records for training and 20% for validation, with the **'train_test_split'** function from the **'scikit-learn'** library.

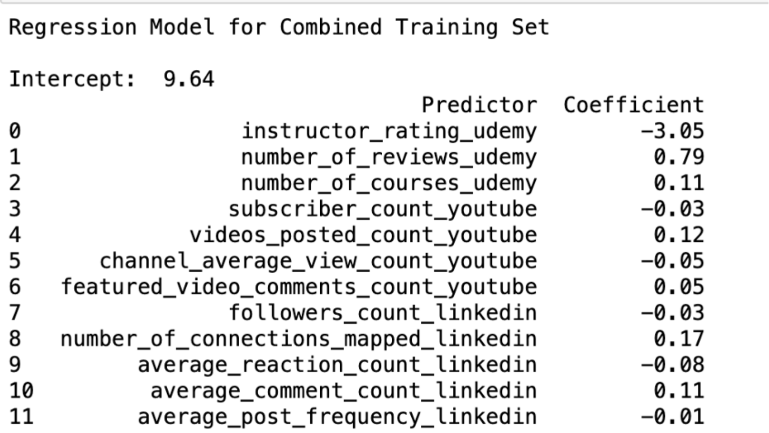
Data Transformation – Log Transformations

To address potential deviations from the normality assumption identified in the assumption checks, where normality was not fully met, log transformations were applied to both predictor variables and the outcome variable.

Log transformation involves taking the natural logarithm of each data point. Log transformation is beneficial in making data more symmetric and stabilizing variance, which can enhance the performance of linear regression models.

Regression Model Development

A multiple linear regression model was developed using the transformed training dataset using LinearRegression(). This involved examining the model's intercept, regression coefficients. Following the model development, predictions were made for both the training and validation datasets using the transformed predictors.



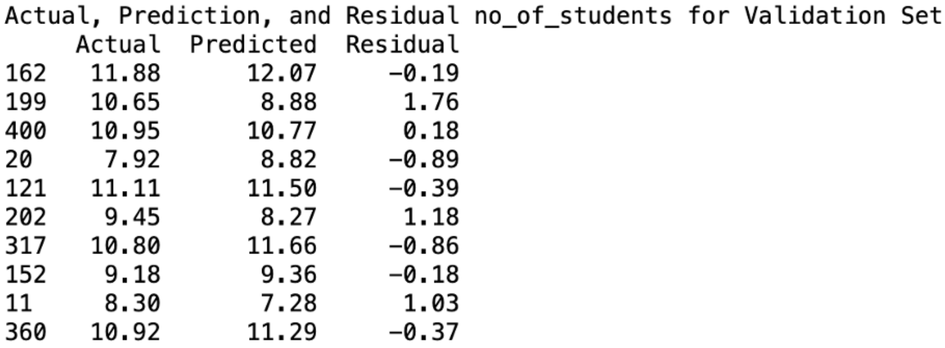
```
Regression Model for Combined Training Set

Intercept:  9.64
                                Predictor  Coefficient
0                 instructor_rating_udemy        -3.05
1                 number_of_reviews_udemy         0.79
2                 number_of_courses_udemy         0.11
3                subscriber_count_youtube        -0.03
4              videos_posted_count_youtube        0.12
5       channel_average_view_count_youtube       -0.05
6    featured_video_comments_count_youtube        0.05
7                  followers_count_linkedin      -0.03
8    number_of_connections_mapped_linkedin        0.17
9          average_reaction_count_linkedin       -0.08
10          average_comment_count_linkedin        0.11
11         average_post_frequency_linkedin       -0.01
```

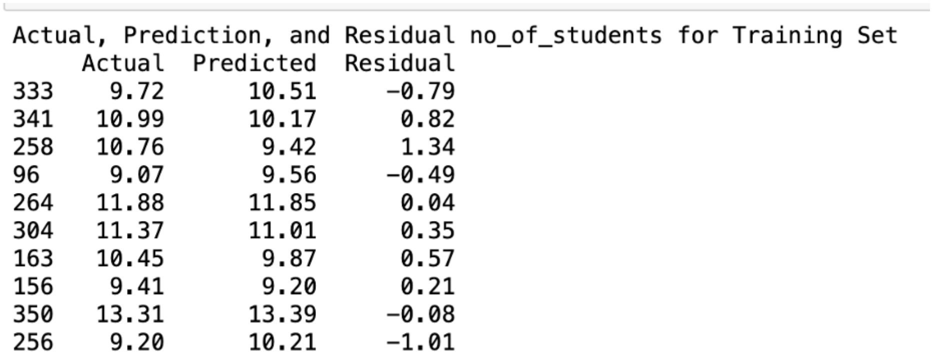
The mathematical equation of this multiple linear regression model is as follows:

number_of_students = 9.64 – 3.05instructor_rating_udemy + 0.79number_of_reviews_udemy + 0.11number_of_courses_udemy – 0.03subscriber_count_youtube + 0.12videos_posted_count_youtube – 0.05channel_average_view_count_youtube + 0.05featured_video_comments_count_youtube – 0.03followers_count_linkedin + 0.17number_of_connections_mapped_linkedin– 0.08average_reaction_count_linkedin + 0.11average_comment_count_linkedin – 0.01average_post_frequency_linkedin

The model's ability to generalize to new, unseen data was assessed by applying it to the validation set. Additionally, predictions were made for the training set to evaluate the model's performance on the data it was trained on.



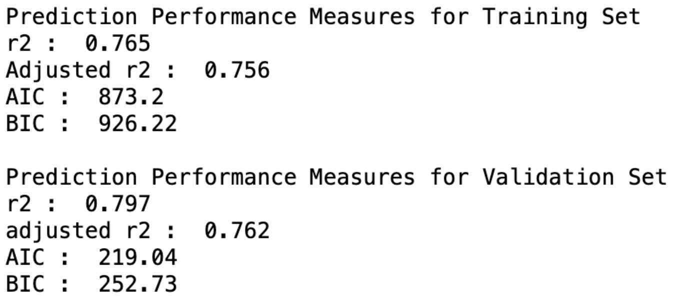
```
Actual, Prediction, and Residual no_of_students for Validation Set
     Actual  Predicted  Residual
162   11.88      12.07     -0.19
199   10.65       8.88      1.76
400   10.95      10.77      0.18
20     7.92       8.82     -0.89
121   11.11      11.50     -0.39
202    9.45       8.27      1.18
317   10.80      11.66     -0.86
152    9.18       9.36     -0.18
11     8.30       7.28      1.03
360   10.92      11.29     -0.37
```



```
Actual, Prediction, and Residual no_of_students for Training Set
     Actual  Predicted  Residual
333    9.72      10.51     -0.79
341   10.99      10.17      0.82
258   10.76       9.42      1.34
96     9.07       9.56     -0.49
264   11.88      11.85      0.04
304   11.37      11.01      0.35
163   10.45       9.87      0.57
156    9.41       9.20      0.21
350   13.31      13.39     -0.08
256    9.20      10.21     -1.01
```

Model Performance Measures

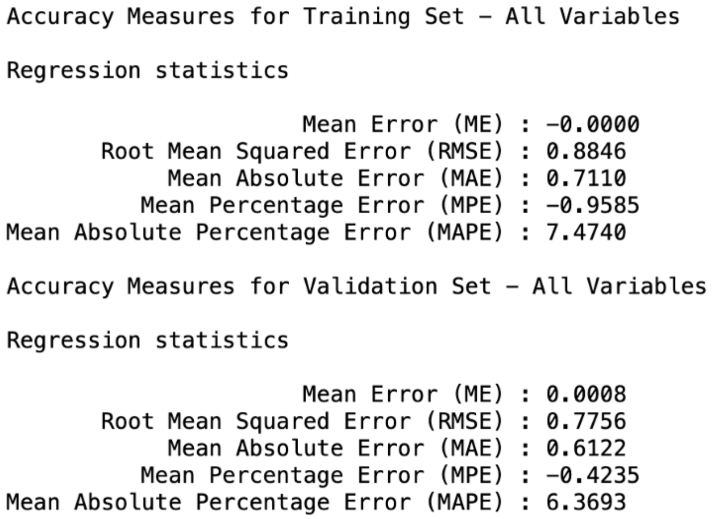
Performance measures, including R-squared, adjusted R-squared, AIC, and BIC, were calculated for both the training and validation sets.



```
Prediction Performance Measures for Training Set
r2 :  0.765
Adjusted r2 :  0.756
AIC :  873.2
BIC :  926.22

Prediction Performance Measures for Validation Set
r2 :  0.797
adjusted r2 :  0.762
AIC :  219.04
BIC :  252.73
```

The difference between the respective R^2 (0.765 for training and 0.797 for validation) and adjusted R^2 (0.756 for training and 0.762) coefficients for training and validation data sets is very small, and therefore, this model does not show any overfitting. At the same time, both R^2 and adjusted R^2 for the validation data set are high and overall indicate a good fit of this multiple regression model to predict number_of_students.



```
Accuracy Measures for Training Set - All Variables

Regression statistics

                      Mean Error (ME) : -0.0000
       Root Mean Squared Error (RMSE) : 0.8846
            Mean Absolute Error (MAE) : 0.7110
          Mean Percentage Error (MPE) : -0.9585
Mean Absolute Percentage Error (MAPE) : 7.4740

Accuracy Measures for Validation Set - All Variables

Regression statistics

                      Mean Error (ME) : 0.0008
       Root Mean Squared Error (RMSE) : 0.7756
            Mean Absolute Error (MAE) : 0.6122
          Mean Percentage Error (MPE) : -0.4235
Mean Absolute Percentage Error (MAPE) : 6.3693
```

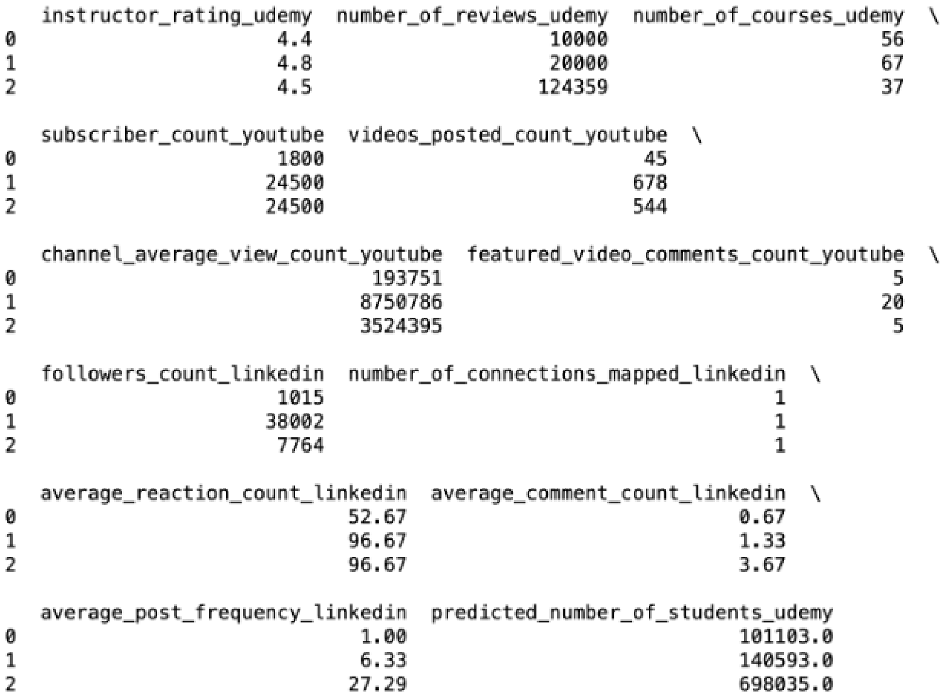
The accuracy measures for the training and validation data sets, particularly RMSE (0.88 vs. 0.77) and MAPE (7.47 vs. 6.36), are almost similar, and thus no overfitting took place. In addition, the validation set's MAPE indicates that the margin of error in predicting number_of_students using the identified regression model is 6.36%.

Predicting New Data

To evaluate the predictive capability of the developed linear regression model, predictions were made for a new dataset using the trained model. The new dataset, represented by the 'new_data' DataFrame, consists of values for all the predictor variables.

These predictor variables, stored in the 'predictors' list, were log-transformed to align with the preprocessing applied during model training. The trained linear regression model ('combined_lm') was then used to make predictions on the transformed data.

To interpret the predictions in the original scale, the predicted values were back-transformed using the inverse of the log transformation. The resulting predictions, along with the original predictor values, are presented in the 'new_predictions_df' DataFrame. This analysis provides insights into the model's performance when applied to new, unseen data.



```
   instructor_rating_udemy  number_of_reviews_udemy  number_of_courses_udemy  \
0                      4.4                    10000                       56   
1                      4.8                    20000                       67   
2                      4.5                   124359                       37   

   subscriber_count_youtube  videos_posted_count_youtube  \
0                      1800                           45   
1                     24500                          678   
2                     24500                          544   

   channel_average_view_count_youtube  featured_video_comments_count_youtube  \
0                              193751                                      5   
1                             8750786                                     20   
2                             3524395                                      5   

   followers_count_linkedin  number_of_connections_mapped_linkedin  \
0                      1015                                      1   
1                     38002                                      1   
2                      7764                                      1   

   average_reaction_count_linkedin  average_comment_count_linkedin  \
0                            52.67                            0.67   
1                            96.67                            1.33   
2                            96.67                            3.67   

   average_post_frequency_linkedin  predicted_number_of_students_udemy  
0                             1.00                            101103.0  
1                             6.33                            140593.0  
2                            27.29                            698035.0  
```

Results and Conclusion

Notable findings include a strong positive correlation (0.886) between the number of Udemy course reviews and the number of students, indicating that courses with more reviews tend to attract a larger audience. On the other hand, relationships with YouTube and LinkedIn engagement metrics exhibit weaker correlations, with values ranging from 0.02 to 0.414. These weaker correlations suggest a less influential connection between these social media metrics and the number of Udemy students.

The transformed predictors and outcome variable improved the model's adherence to regression assumptions.

Step 6: Data Visualization

The aim behind the dashboards was to uncover patterns followed by existing online educators so that John Doe’s strategy for being an online educator can be driven by proven patterns for

success. The primary question that drove each decision in making, analyzing and retaining visualizations was “What should John Doe do to be successful as an online educator? What has worked so far, what will work for him?”

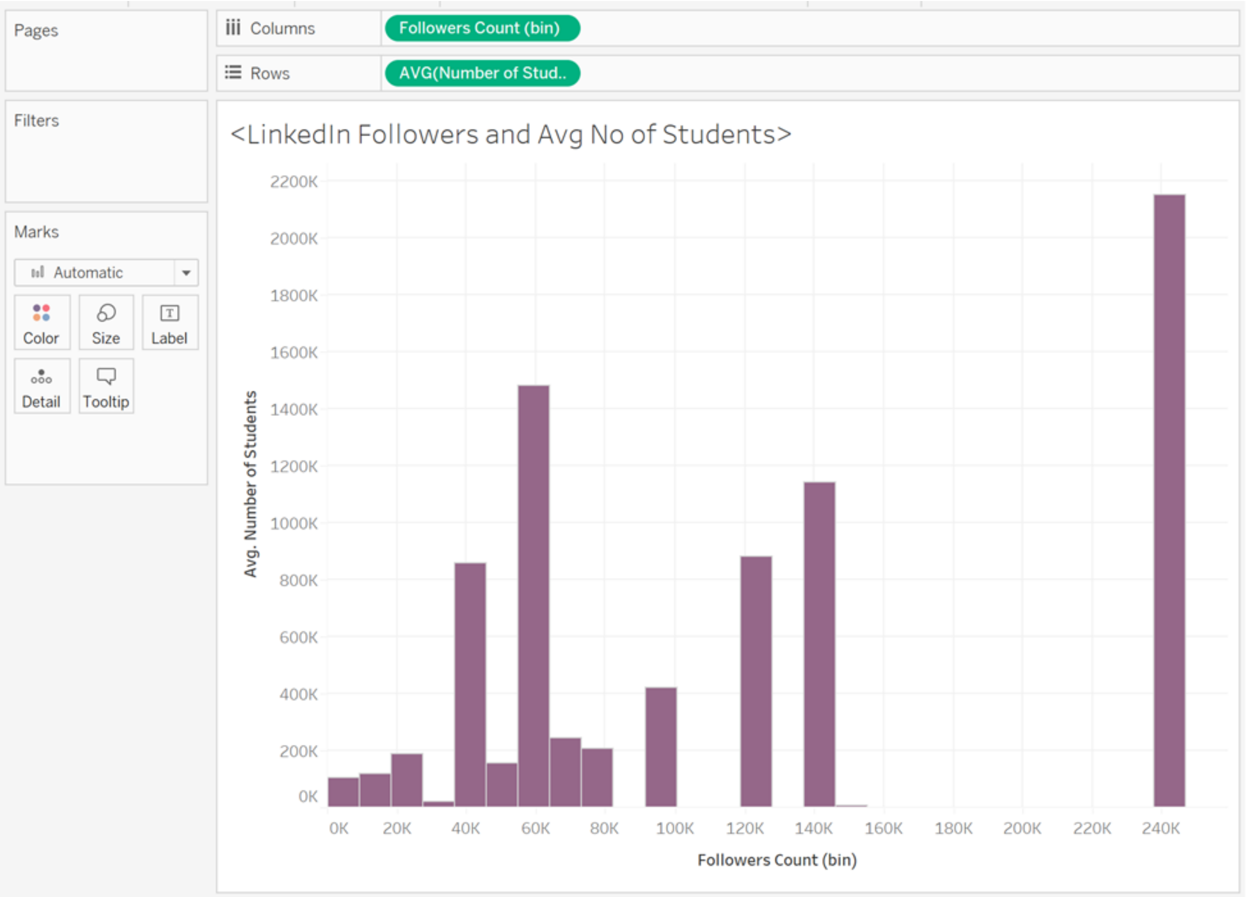
Tableau served as the visualization tool. The result of the data cleaning process using AWS Glue in step 4 of the report was used as input for the AWS Athena database. The Athena database was then organized into tables, including udemy_data, youtube_data, and linkedin_data. These tables were joined to create combined_data, retaining records of educators with Udemy, LinkedIn, and YouTube pages through an inner join. A subsequent table, combinedudemy_top50, identified the top 50 online educators based on Udemy student count.

Two sets of visualizations were created: one exploring the relationship between Udemy and LinkedIn, and another exploring Udemy and YouTube. In the first visualization, an inner join was performed on Tableau between udemy_data and linked_data, resulting in 457 records of online educators present on both platforms. Key variables imported included instructor rating, number of courses, number of reviews, number of students (Udemy data) and average comment count, average post frequency, average reaction count, followers count, number of connections mapped (LinkedIn data). The second visualization involved a similar inner join on udemy_data and youtube_data, producing a table with 727 records. This introduced new variables for YouTube data – namely channel average view, featured video comments, subscriber count, videos posted – alongside the existing Udemy variables.

Udemy and LinkedIn Dashboard

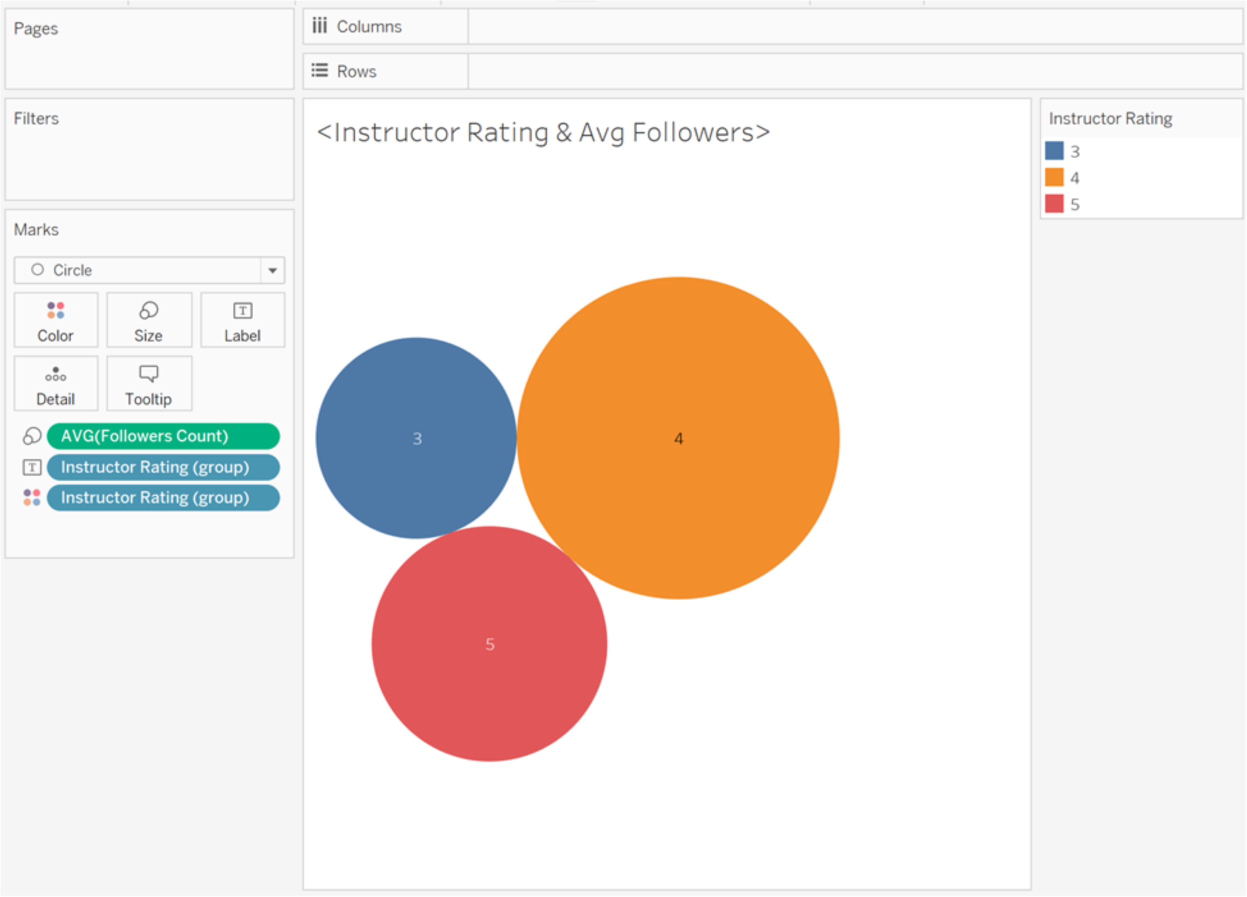
Having five subplots, the dashboard begins with an overview of the data – number of records, and minimum and maximum number of students – to show that the data is not normally distributed.

The first graph explores the relationship between number of followers on LinkedIn and number of students on Udemy. These were the only two variables that showed moderate correlation between the two platforms during the initial stages of data analysis. That correlation shows through in the bar chart.



For this chart, the number of followers variable on linkedin_data table was binned using Tableau. This new variable was placed along the x-axis with average number of students along the y-axis. The overall trend is that as the number of followers increases, the number of students on udemy also increases.

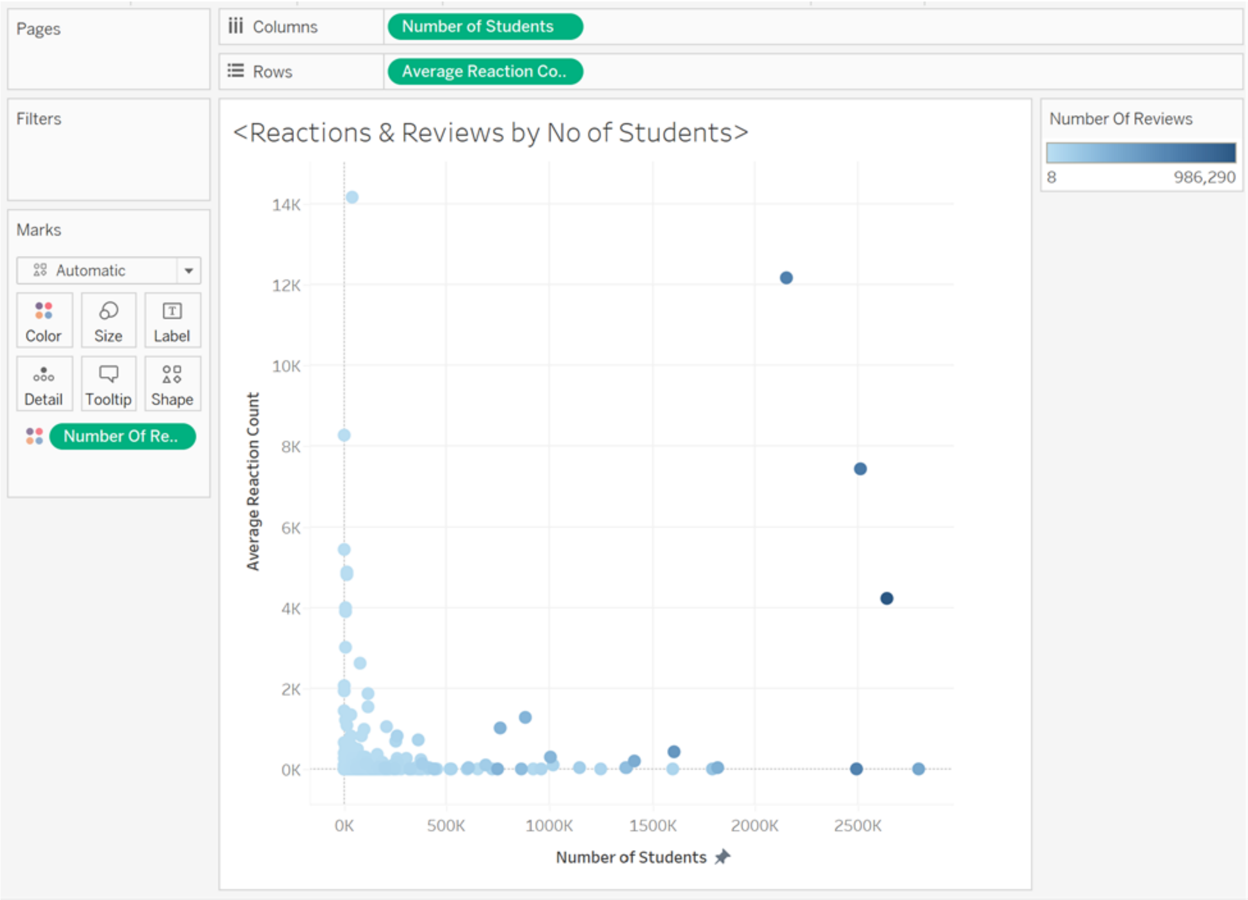
The second visualization is a packed bubble chart that shows the instructor rating on udemy and average number of followers for each rating group. To create this chart, instructor rating variable was grouped on Tableau, which became a new variable. Here the size of the bubbles represents average followers on LinkedIn.



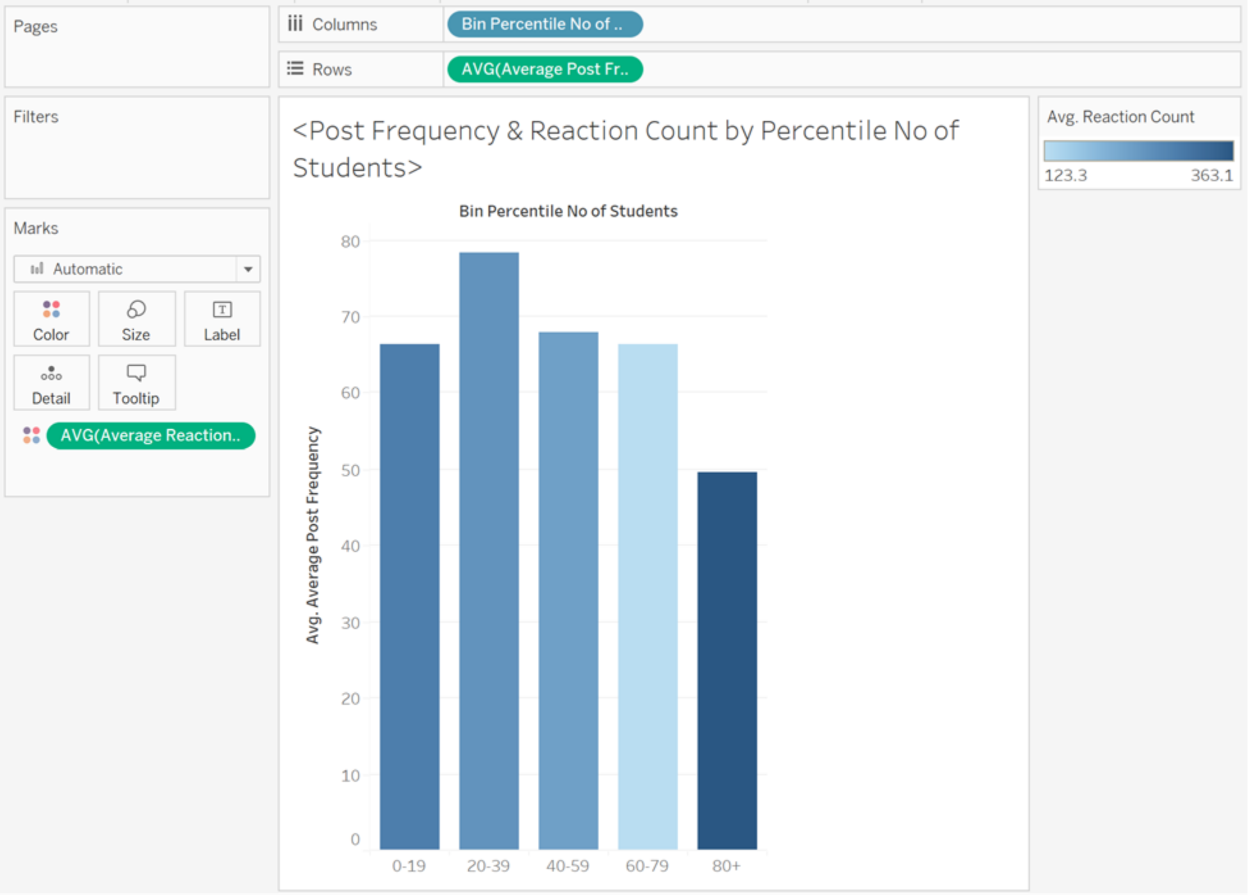
AVG(Followers Count)
Instructor Rating (group)
Instructor Rating (group)

It could be that since it's difficult to get a perfect 5/5 rating on Udemy, especially as number of students increases, the highest average number of followers is recorded by those educators that have a 4/5 rating.

The third chart is a scatterplot that explores the relationship between number of reactions received on LinkedIn posts and number of students on Udemy. Number of students is plotted along the x-axis, and average reaction count along the y-axis, both unaggregated to show all 457 records individually. In addition, the color of each pointer indicates the number of reviews on Udemy – darker the pointer, more the number of reviews. The chart does not show any significant relationship or pattern among the variables.



The final visualization in this dashboard is a bar graph that represents the patterns among post frequency, reaction count and percentile number of students. To create this graph, the first variable that had to be created was the percentile of number of students. For each value in the number of students column of the udemy_data table, the corresponding percentile rank of that variable was calculated. Then 5 bins were created using those percentiles – those records where number of students was between 0-19th percentile, 20-39th percentile, 40-59th percentile, 60-79th percentile and those records where number of students was equal to 80th percentile or more. This new variable was called Bin Percentile No of Students, and was plotted along the x-axis. Along the y-axis was average post frequency for each of these bins. Additionally, the color of the bars represented average reaction count recorded on LinkedIn.



Columns: Bin Percentile No of ..

Rows: AVG(Average Post Fr..

Marks: Automatic

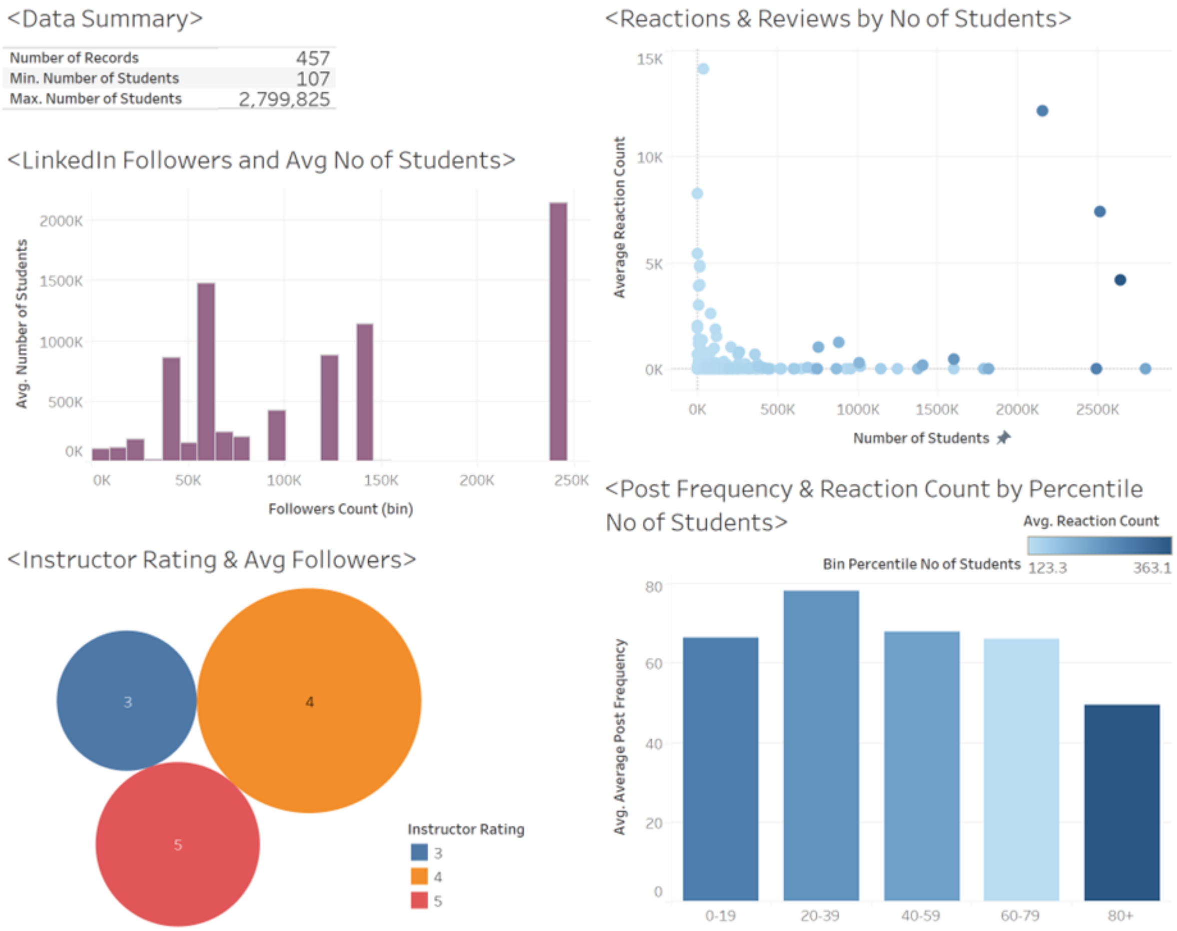
AVG(Average Reaction..

Pages
Filters
Marks
Color
Size
Label
Detail
Tooltip
<Post Frequency & Reaction Count by Percentile No of Students>
Bin Percentile No of Students
Avg. Average Post Frequency
80
70
60
50
40
30
20
10
0
0-19
20-39
40-59
60-79
80+
Avg. Reaction Count
123.3
363.1

According to the patterns in the graph, we see that those online educators who have relatively low number of students – in the 0-19, 20-39 percentile bins – receive relatively high engagement on LinkedIn despite a lot of time between posts on LinkedIn – 60+ days.

Additionally, we see that when online educators break a threshold and attain massive growth on Udemy (marked by over 80th percentile number of students), they may choose platforms like LinkedIn to keep in touch with their large audience – more frequent posts, and higher engagement metrics.

The final dashboard with all the above subplots looks like this:



<Data Summary>

Number of Records	457
Min. Number of Students	107
Max. Number of Students	2,799,825

<Reactions & Reviews by No of Students>

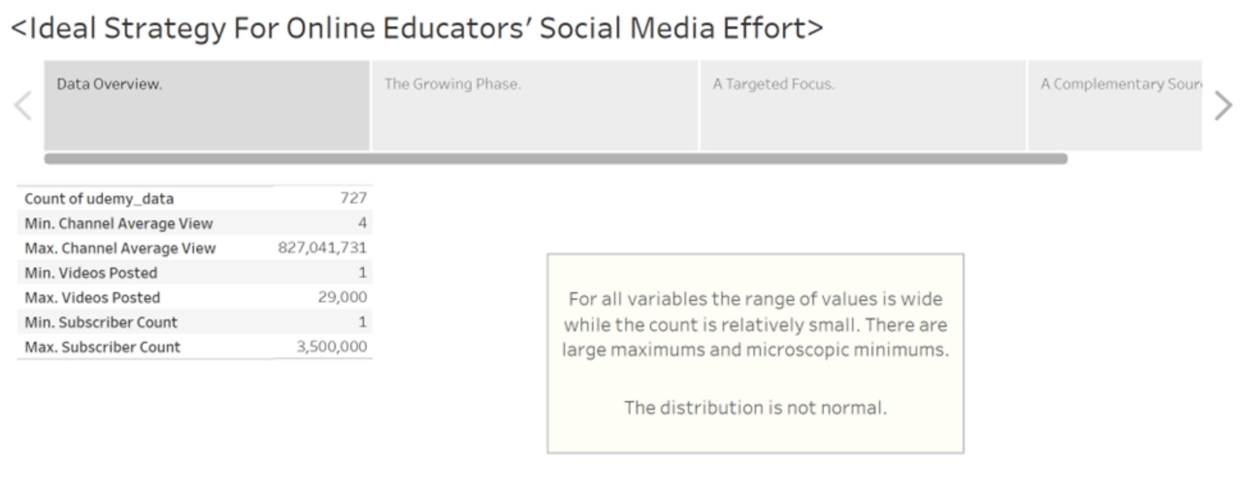
<LinkedIn Followers and Avg No of Students>

<Post Frequency & Reaction Count by Percentile No of Students>

<Instructor Rating & Avg Followers>

Udemy and YouTube Visualizations

A bunch of patterns were uncovered in the visualizations that explored Udemy and Youtube, all of which were organized into a Tableau Story. Each page of the story is explored below in detail, starting with the data overview. Similar to the overview in Udemy and LinkedIn dashboard, this too shows the number of records and maximum and minimum values for number of subscribers, videos of posted and channel average view count.



<Ideal Strategy For Online Educators' Social Media Effort>

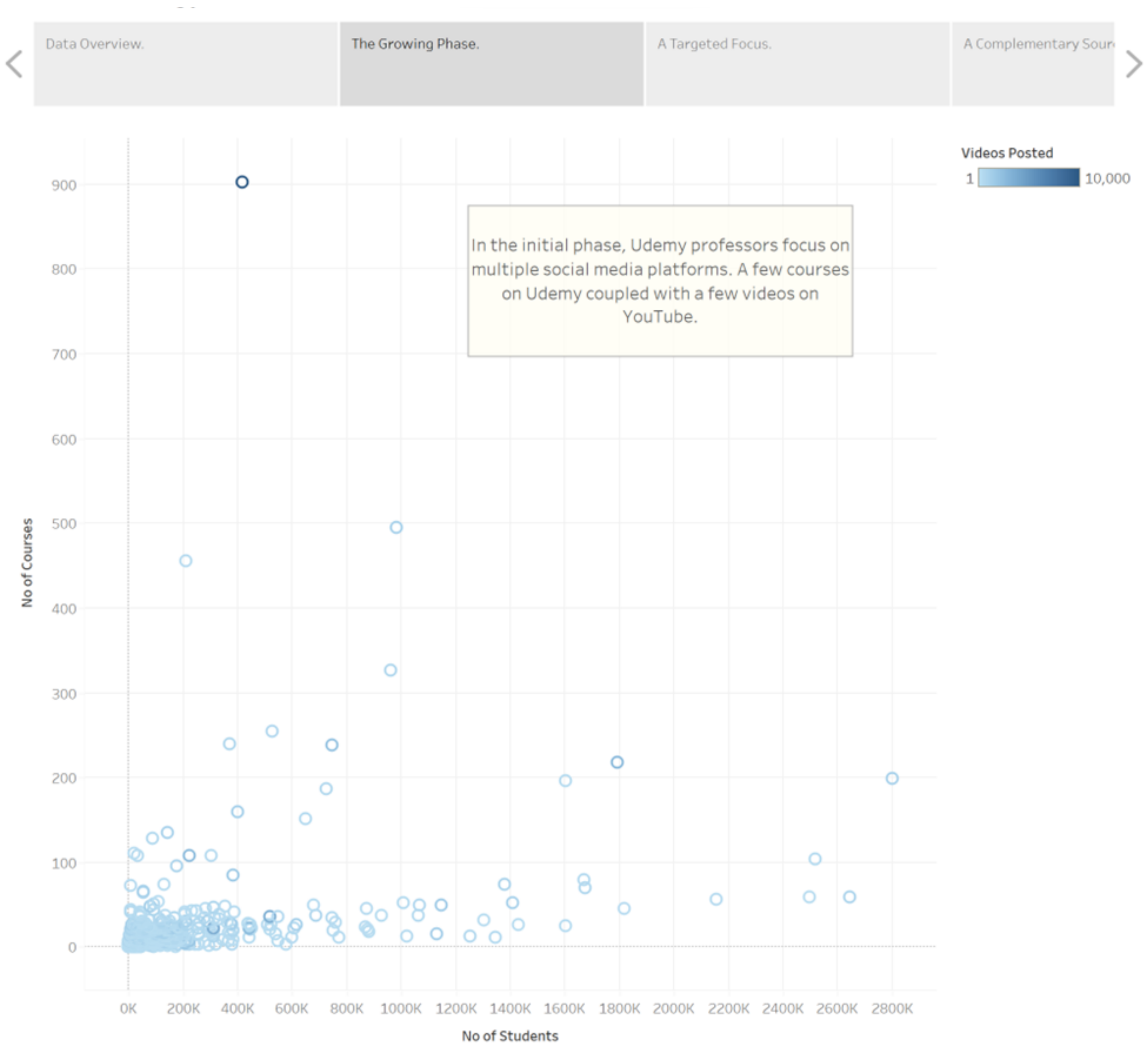
Count of udemy_data	727
Min. Channel Average View	4
Max. Channel Average View	827,041,731
Min. Videos Posted	1
Max. Videos Posted	29,000
Min. Subscriber Count	1
Max. Subscriber Count	3,500,000

For all variables the range of values is wide while the count is relatively small. There are large maximums and microscopic minimums.

The distribution is not normal.

In this page – The Growing Phase – we would like to focus on the cluster of records near the origin, referring to the people who have just begun their journey as online educators. Number of students was plotted along the x-axis, number of courses on Udemy along the y-axis, with the color of the pointer referring to the number of videos on YouTube.

As people begin their efforts on online platforms, we see that they post a lot of videos on YouTube, create courses on Udemy and their number of students is relatively low, but steadily increasing. A predominant number of educators are active on both Udemy and YouTube in their starting stages. This seems to help a small portion of educators who enjoy growth on both platforms (marked by pointers along a slope).



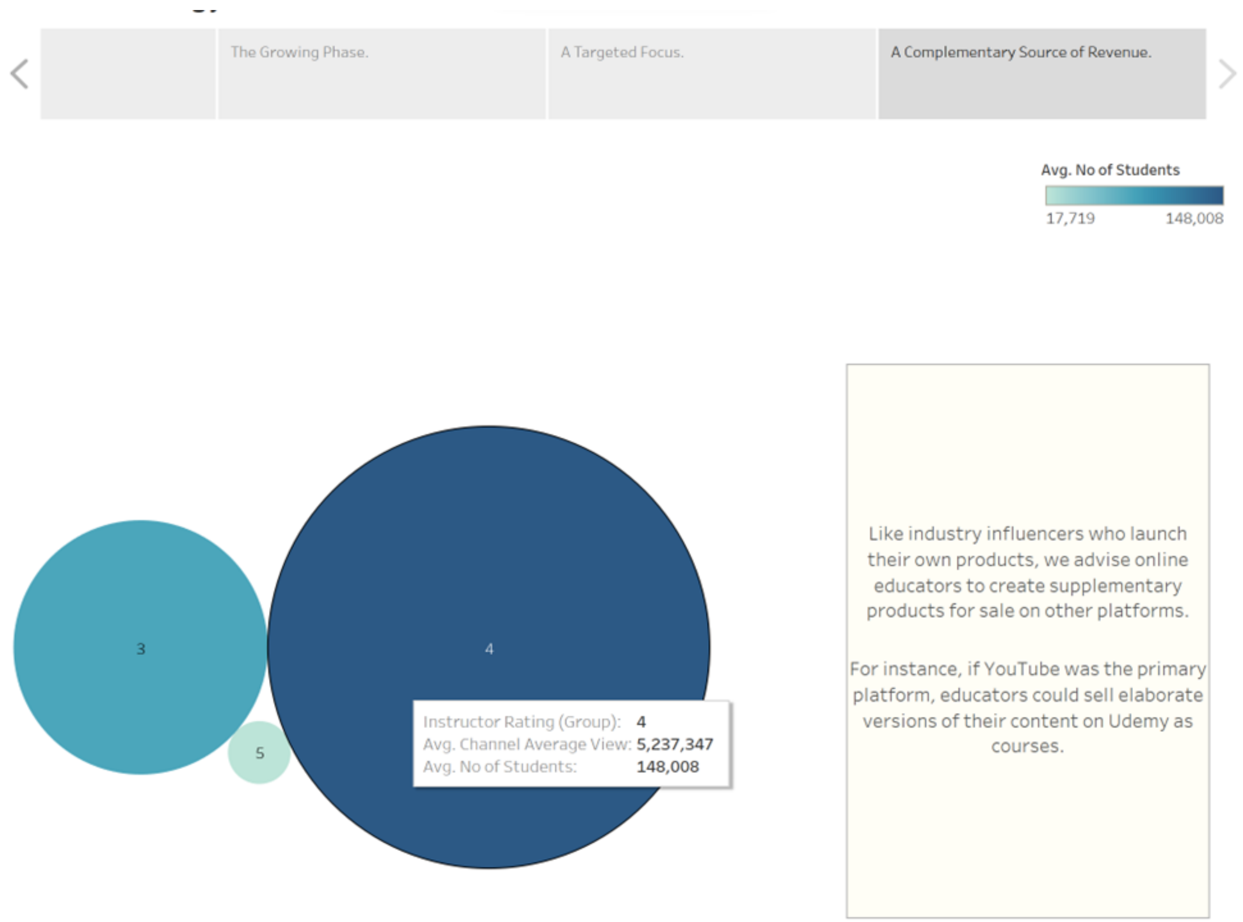
In the initial phase, Udemy professors focus on multiple social media platforms. A few courses on Udemy coupled with a few videos on YouTube.

In the third page, what we called “A Targeted Focus”, we see that most records are along an axis, but almost never along a slope. To create this chart, number of students remained on the x-axis, number of subscribers was plotted along the y-axis, and the color of each pointer was set to indicate instructor rating. It is important to note that in all 727 records visualized here, less than 10 had a perfect instructor rating of 5/5.

What does it mean if the records are along an axis but rarely along a slope? Once the initial growth phase is over and educators receive a certain amount of engagement on one platform, they divert the majority of their efforts on that platform. Take for example those records along the x-axis. Those are the online educators whose courses have over a million students on Udemy, they have received 4/5 instructor ratings (which is a difficult feat), but their number of videos on YouTube is very less.

Similarly, those records along the y-axis have hundreds of thousands of videos on YouTube, and lower hundred-thousands of students on Udemy, and an incredible 4/5 instructor rating.

This leads us to the final page of the Tableau Story, “A Complementary Source of Revenue”. In addition to focusing on one platform, most educators also leverage a secondary platform where they make money. The fact that YouTube educators have hundreds of thousands of students of Udemy and 4/5 instructor rating indicates that they may sell elaborate versions of their courses on Udemy for a supplementary source of income.



Instructor Rating (Group): 4
Avg. Channel Average View: 5,237,347
Avg. No of Students: 148,008

Like industry influencers who launch their own products, we advise online educators to create supplementary products for sale on other platforms.

For instance, if YouTube was the primary platform, educators could sell elaborate versions of their content on Udemy as courses.

In conclusion, we recommend that John Doe follow a similar strategy as well. Start off by posting on multiple platforms and creating an audience. Then as engagement takes off in one platform, focus primarily on that while keeping the others as secondary platforms for revenue. It’s important to create products (in this case courses on Udemy or playlists on YouTube) in the secondary platform that can be utilized for cross-platform promotion. Additionally, once John Doe crosses a certain threshold, for communicating and engaging with the audience he may use LinkedIn or such social media platforms.

Results and Business Recommendations

Topic modeling on LinkedIn data offers insights into users' professional interests and expertise. The identified topics are coherent and provide valuable information for networking, and industry trend analysis. From the individual records analyzed, “passionate” “teach” had a relatively high probability of occurrence. However, an interesting insight was that only 85 out of 457 instructors had mentioned Udemy on their LinkedIn bios. Despite that, none of the 85 had actively promoted their Udemy page on LinkedIn - as is evident by the topic modeling algorithm failing to pick up any topics where the word “Udemy” had the highest probability.

The regression model, despite its predictive performance, indicates that the relationship between social media engagement and Udemy course popularity is complex. While the model suggests certain predictors influence course popularity positively (coefficients > 0) and others negatively (coefficients < 0), the overall interpretation should consider the context of the specific data and the characteristics of the instructor. It's important to note that correlation does not imply causation, and other factors beyond the analyzed variables may contribute to course popularity.

No consistent link was found between LinkedIn activity and Udemy success, except for high Udemy student counts correlating with a big LinkedIn following. Successful Udemy instructors were often found using LinkedIn to stay connected once they hit the 80th percentile in student numbers.

On YouTube, we recommend educators to begin by being active on various platforms and subsequently narrowing down on the one gaining traction. We recommend the use of secondary platforms for extra income by creating cross-platform content like Udemy courses or YouTube playlists. As success grows, platforms like LinkedIn become key for engaging with the large audience.

Business Recommendations

Based on the insights derived from this comprehensive analysis of Udemy, LinkedIn, and YouTube data, we propose the following business recommendations for John Doe to improve the chances of his Udemy success:

Leverage LinkedIn for Professional Branding

John should strategically use LinkedIn as a powerful tool for professional branding. The platform's evolution into a content-sharing hub provides educators with an opportunity to showcase expertise, share insights, and directly engage with their audience. By consistently posting high-quality content related to his courses, he can establish a strong professional presence and potentially attract a broader audience.

Strategic Use of Platforms for Content Monetization

YouTube's shift towards enabling content creators to monetize their videos presents an opportunity for Udemy instructors like John to diversify their income streams. John is advised to initiate his online presence on multiple platforms and gauge traction. Once a platform gains momentum, focusing on content creation for monetization on that platform, such as Udemy course snippets or exclusive content, can contribute to additional revenue.

Cross-Platform Promotion for Audience Expansion

To maximize audience reach and engagement, instructors should adopt a cross-platform promotion strategy. Creating content, such as Udemy courses or YouTube playlists, that can be utilized for promotion across different platforms can amplify the visibility of an John’s expertise. This approach allows for a more diversified audience base and increased opportunities for course enrollment.

By implementing these recommendations, John Doe can optimize his online presence, engage a broader audience, and potentially boost the popularity of their Udemy courses.

Further Improvements

Taking this project further, we can explore/analyze YouTube video description section where we have observed Udemy links being promoted by instructors. Analyzing that data could give us additional insights.

If a correlation between engagement activities on LinkedIn and YouTube and Udemy course popularity is established, our investigation will delve deeper. We will analyze the types of content instructors are posting on these platforms and aim to identify optimal strategies. By segregating professors with a strong LinkedIn or YouTube presence, we aim to guide instructors like John Doe on content themes, delivery style, speech pace, and more. This nuanced approach will help instructors not only attract learners but also provide them with engaging and impactful educational experiences.

We suspect that the lack of data about post impressions on LinkedIn severely limited the insights that could be drawn about the relationship between LinkedIn and Udemy success. In taking this further, we would recommend exploring ways to capture additional information about LinkedIn engagement metrics like impressions to add the much needed data points.

Summary of Tools Mentioned

- Python Language is used to conduct web scraping,data analysis and text mining.
- Amazon S3 (Simple Storage Service) is a scalable cloud storage service provided by Amazon Web Services (AWS) designed to store and retrieve data in the form of objects.
- AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data from various sources for analytics and data processing.
- AWS Glue Crawler is an automated tool that scans and analyzes data sources such as databases, S3 buckets, and other storage repositories to discover schema and metadata information, facilitating efficient data cataloging and organization.
- AWS Athena is an interactive query service that allows you to analyze data stored in Amazon S3 using standard SQL queries, enabling you to gain insights quickly and easily from your data without the need to set up and manage complex data infrastructure.

- Tableau is a popular business intelligence (BI) and data visualization tool used to transform raw data into interactive and insightful visual representations, aiding in decision-making and analysis.

Contributions

Step	Person	ChatGPT's Inputs
Data Collection	Neenu	Debugging, methods & functions, article
Data Cleaning	Ruchika	Not very helpful for AWS Glue, article
Text Mining	Apurva	Interpretation of results, article
Regression	Ruchika & Neenu	Transformation methods, code (mostly unhelpful), debugging, interpretation
Data Visualization	Swetha	Article, idea generation

References

6.3. Preprocessing Data. scikit. (n.d.). https://scikit-learn.org/stable/modules/preprocessing.html#non-linear-transformation

Akinyemi, A. A. (2023, February 22). Factors That Promote Student Retention in MOOCs | UCI Division of Teaching Excellence and Innovation. UCI DTEI. Retrieved August 30, 2023, from https://dtei.uci.edu/2023/02/22/factors-that-promote-student-retention-in-moocs/

Analysis of Udemy Subjects & Courses. (n.d.). Tableau Software. Retrieved September 11, 2023, from https://public.tableau.com/views/AnalysisofUdemySubjectsCourses/Dashboard1?%3Adisplay_static_image=y&%3AbootstrapWhenNotified=true&%3Aembed=true&%3Alanguage=en-US&:embed=y&:showVizHome=n&:apiID=host0#navType=0&navSrc=Parse

Bali, E. (n.d.). Udemy. Wikipedia. Retrieved August 30, 2023, from https://en.wikipedia.org/wiki/Udemy

Bankoff, C. (2023, June 15). Social Media Marketing Guide for Educational Institutions. Constant Contact. Retrieved August 30, 2023, from https://www.constantcontact.com/blog/social-media-marketing-guide-for-educational-institutions/

Chaudhry, R. (2023, June 14). 8 Effective Strategies to Promote Online Courses Using Social Media. Think Orion. Retrieved August 30, 2023, from https://www.thinkorion.com/blog/how-to-promote-online-course-on-social-media

Du, H. (2020). *Analytics for Udemy courses—Python* [Jupyter Notebook]. https://github.com/peterdu98/udemy-courses-analytics

Faculty Voice: Enhancing The Course Experience Using Social Media | Digital Learning & Innovation. (n.d.). Retrieved August 30, 2023, from https://www.bu.edu/dli/2021/05/17/faculty-voice-enhancing-the-course-experience-using-social-media/

Giri, Y. (2020, September 30). Udemy Courses Data Exploration. *Medium*. https://medium.com/@madeyudhagiri/udemy-courses-data-exploration-8673b68c9548

guide, s. (n.d.). LinkedIn. Wikipedia. Retrieved August 30, 2023, from https://en.wikipedia.org/wiki/LinkedIn

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. *Proceedings of the First ACM Conference on Learning @ Scale Conference*, 41–50. https://doi.org/10.1145/2556325.2566239

JCharisTech (Director). (2021, February 10). *Data Science Project—Exploratory Data Analysis (Udemy Dataset)*. https://www.youtube.com/watch?v=7IAfs5i03ZM

Lei, Z. (2021). *How Do We Design a Good Online Course for Business Analytics?* [HTML]. https://github.com/leizhipeng/analyze_mooc_course_data (Original work published 2021)

McLachlan, S. (2023, April 6). 85+ Important Social Media Advertising Stats to Know. Hootsuite Blog. Retrieved August 30, 2023, from https://blog.hootsuite.com/social-media-advertising-stats/

Notermans, M. (2023, June 14). 14 Easy Ways To Promote An Online Course in 2023. Think Orion. Retrieved August 30, 2023, from https://www.thinkorion.com/blog/how-to-promote-an-online-course

RPubs—Udemy Online course Analysis. (n.d.). Retrieved September 11, 2023, from https://rpubs.com/nnaemeka/udemy

Shah, C. (2023, April 2). Exploring Udemy Courses Trends Using Google Big Query. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2023/04/exploring-udemy-courses-trends-and-insights-with-google-big-query/

Udemy Courses. (n.d.). Retrieved September 11, 2023, from https://www.kaggle.com/datasets/andrewmvd/udemy-courses

Udemy Courses Data 2023. (n.d.). Retrieved September 11, 2023, from https://www.kaggle.com/datasets/ankushbisht005/udemy-courses-data-2023

Udemy Courses Data Analysis—Notebook by Abishek V (abishekv99) | Jovian. (n.d.). Retrieved September 11, 2023, from https://jovian.com/abishekv99/udemy-courses-data-analysis

Wikipedia. (n.d.). YouTube. Wikipedia. Retrieved August 30, 2023, from https://en.wikipedia.org/wiki/YouTubeCrane, R. A., & Comley, S. (2021). Influence of social learning on the completion rate of massive online open courses. *Education and Information Technologies*, *26*(2), 2285–2293. https://doi.org/10.1007/s10639-020-10362-6

Yang, S., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2022). The science of YouTube: What factors influence user engagement with online science videos? *PLOS ONE*, *17*(5), e0267697. https://doi.org/10.1371/journal.pone.0267697

Zheng, S., Han, K., Rosson, M. B., & Carroll, J. M. (2016). The Role of Social Media in MOOCs: How to Use Social Media to Enhance Student Retention. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, 419–428. https://doi.org/10.1145/2876034.2876047

Appendix A - Data Dictionary

Udemy data:

index_number - unique identifier for each instructor

udemy_course_url – url of Udemy course

instructor_url – url of instructor data

instructor_rating – instructors rating in Udemy

number_of_students – number of students enrolled under instructor

number_of_courses - number of courses taken by instructor so far

linkedin_url – instructors LinkedIn page url

youtube_url – instructors YouTube page url

YouTube data:

index_number: unique identifier for each instructor

channel_title: title of instructors YouTube channel

subscriber_count: number of subscribers

number_of_videos_posted: number of videos posted by instructor so far

channel_created_date: creation date of the youtube channel

channel_average_views_count: average number of views got for the channel

channel_description: description given by the instructor

number_of_comments_for_featured_video: number of comments got for the featured video

featured_video_comments: top 20 comments of the featured video

LinkedIn data: (Personal page)

index_number: unique identifier for each instructor

followers_count_linkedin: number of followers of the page

number of connections: number of connections will be 1 if person has more than 500 connections and 0 for less than 500

average_reaction_count: average count of the number of reactions got for latest 3 posts in page

average_comment_count: average count of the number of comments got for latest 3 posts in page

average_post_frequency: average count of the frequency between posts for latest 3 posts in page

Appendix B - Data, Code, Articles

The following code and data files are on a shared GitHub repository.

1. The code files are:
    a. WebScraping_Udemy_(data1).ipynb
    b. WebScraping_Udemy_(data2).ipynb
    c. LinkedInScraping.ipynb
    d. Youtube_Scraping.ipynb
    e. udemy-datacleaning-recipe.json
    f. youtube-datacleaning-recipe.json
    g. linkedin-instructor-datacleaning-recipe.json
    h. Sentiment_Analysis_On_Youtube_Data.ipynb
    i. Topic_modelling_On_LinkedIn_data.ipynb
    j. multiple_linear_regression.ipynb
    k. Udemy_LinkedIn_VGit.twb
    l. Udemy_YouTube_VGit.twb
2. Initial Kaggle datasets - 5000 and 3000 links to Udemy courses.
3. Udemy, YouTube and LinkedIn data from scraping the above files .
    a. combined_udemy_data.csv - raw data
    b. combined_youtube_data.csv - raw data
    c. LinkedIn_instructor_data.csv - raw data
    d. LinkedIn_company_data.csv - raw data
    e. youtube_regression.csv - cleaned data for regression
    f. linkedin-instructor_regression.csv - cleaned data for regression
    g. udemy_regression.csv - cleaned data for regression
    h. Youtube_data_comments_segregated.csv - cleaned data for text mining
    i. udemy-datacleaning-_2023_10_16.csv - cleaned data for AWS Athena
    j. youtube-datacleaning_2023_10_16.csv - cleaned data for AWS Athena
    k. linkedin-instructor-datacleaning-_2023_10_16.csv - cleaned data for AWS Athena

Articles published on different sections of the report are as follows.

1. Unleashing the Power of Web Scraping for Udemy, YouTube and LinkedIn

2. Impact of Sentiment Analysis - Navigating the Landscape of YouTube Comments
3. Power of Topic Modeling - Crafting a Compelling LinkedIn About Section
4. Unleashing the Potential of Data Transformations with AWS Glue Databrew
5. LinkedIn post data visualization