Lucerne University of Applied Science and Arts

Master of Science in Applied Information and Data Science

Computer Science Concepts for Data Scientists

Udemy Course Analysis

(Information Visualization)

submitted to

Dr. Halldor Janetzko

Prof. Dr. Martin Zimmermann

by

Quyen Duong

Submission deadline: June 13th, 2021

Table of contents

List of figures

List of tables

1. Motivation

Online learning has recently become one of the common options not only for students but also for those who would like to broaden their knowledge and gain new skills in specific domains. Because of the significant benefits such as accessibility and flexibility, and especially after the pandemic, online learning can be the future and likely take the place of land-based learning in a certain way. Over the last decade, numerous online courses have been created. To name but a few, Udemy, Coursera, Lynda, Skillshare, Udacity are the most popular online learning platforms that serve millions of users all over the world.

This analysis report is inspired by Udemy, one of the top-trending online platforms, offering more than 155 000 courses by 56 000 instructors for more than 40 million learners in more than 65 languages up to date (Udemy, 2021). Udemy, founded in May 2010, is an American online learning platform aimed at professional adults and students that offers both free and paid courses. Anybody can create an online class. This is the business model which grants Udemy to have diverse lessons. Students take courses vastly to improve life and job-related skills: some courses generate credit toward technical certifications. Udemy has made a special effort to attract corporate trainers seeking to create coursework for employees of their company.

Using the dataset found on Kaggle[1] about courses on Udemy, we analyze the performance of online courses and getting insights about users' preferences in terms of course price, level, course duration. This analysis report is carried out by several main steps: data cleaning, data visualization, and interpretation of the findings.

2. Research questions

Numerous questions can be initiated from the dataset. However, the project only concentrates on the popularity of Udemy courses from the dataset. Deriving from the motivation, the research questions are formulated as below:

- What are the most popular courses and subjects in Udemy?
- Which factors influence the course popularity within Udemy?

Firstly, it is essential to clarify the term popularity based on the dataset as the state of being subscribed by abundant people. In other words, I can measure the popularity level by ranking the courses and subjects regarding their subscriber number. The number of reviews can also be relevant, but I will focus more on subscribers as a key variable in our analysis. The second question is an open question that allows exploring the data by associating subscribers with other variables such as subjects, content duration, free or paid courses, price, and course levels. By doing so, it enables us to investigate and discover exciting trends as well as patterns that the data display. To answer these questions, I attempt to tell the story and analysis results through data visualization. Meaningful graphs will be presented in a dynamic dashboard.

3. Methodology

The methodology of this project consists of several steps that could be sequential or iterative depending on the tasks. The authors implement an iterative process rather than solely waterfall, especially from step 3 to step 8, to generate a worthwhile dashboard and not omit important information.

- Step 1: Obtaining and learning the data features

[1] Dataset on https://www.kaggle.com/andrewmvd/udemy-courses (Assessed date: 16th April 2021)

- Step 2: Cleaning and processing the data
- Step 3: Defining the research questions
- Step 4: Starting with basic visualizations. Diagnosing the messages from basic visualizations, forming the most illuminating indicators
- Step 5: Choosing the appropriate chart types
- Step 6: Designing color, fonts, scales, legends, size, shapes, labels to illustrate critical messages to the audiences effectively
- Step 7: Creating the dashboard that serves the research questions
- Step 8: Describing, analyzing, and discussing the findings from the dashboard

4. Target audience

First and foremost, the analysis and dashboard can be an excellent reference for Udemy company about the overview of their course performance and then make decisions based on data-driven visualization regarding sales, marketing, or customer relationship management. Second, course developers can fully track popular courses or subjects' critical determinants for course choices to create more valuable and practical content. Finally, learners are often confused and do not know what they should learn. Hence, the analysis could bring them the idea to choose appropriate courses based on subjects, duration, price, subscribers.

5. Data description

In this project, I look at various courses offered by Udemy on the available data from the publishing year 2011 to 2017. Noticeably, it only covers the first six months of 2017. There are 3,682 observations (rows) and 12 features (columns). Udemy courses are divided into four main subjects (business finance, graphic design, musical instruments, and web design) (Larxel, 2020).

No.	Column name	Data type	Description
1	course_id	Integer	Course ID
2	course_title	String	Course title
3	url	String	Course URL
4	is_paid	Boolean	Whether the course is free or paid
5	price	Integer	Course price
6	num_subscribers	Integer	Number of subscribers
7	num_reviews	Integer	Number of reviews
8	num_lectures	Integer	Number of lectures
9	level	Object	Course levels
10	content_duration	Float	Duration of the course material
11	published_timestamp	String	The date that the course was published
12	subject	String	Course subject

Table 1: Dataset description

6. Data cleaning and processing

6.1 Drop the irrelevant columns

I drop the "course_id" column by using "cut", which is commonly employed for extracting columns from a file based on their indices (anticipated by "-f"). These indices are given by the position of text in each line concerning the delimiter of the file. However, thanks to the "--complement" option, I implement the opposite in our case. Thus, I *exclude* the columns marked by the given indices from the file. The syntax is, therefore:

$ cut --complement –f [indices] [filename]

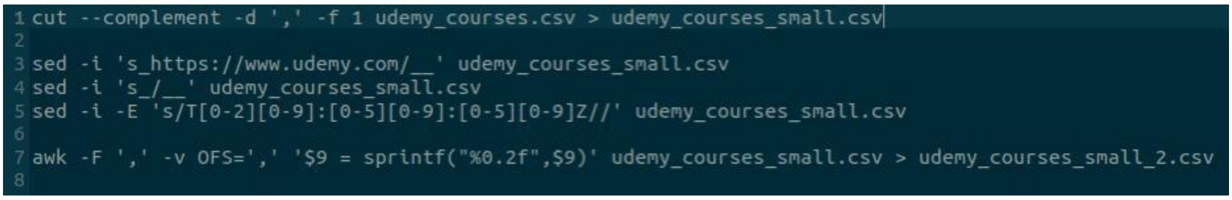
6.2 Convert the "published_timestamp" column to DateTime format

Here I remove unwanted characters from strings with "sed". For example, I remove the publishing time of a course (hours, minutes, seconds) that follows the corresponding date (Figure 1). The syntax of the employed "sed" command (omitting other options, which I do not explain here) is:

$ sed 's/[text_to_find]/[text_to_replace]/' [filename]

6.3 Round the number in the "content_duration" column to two decimals

For this purpose, I exploit "awk" to apply "sprintf" to the values in the "content_duration" column. The full command is quite complicated and can be seen below (line 7).



```
1 cut --complement -d ',' -f 1 udemy_courses.csv > udemy_courses_small.csv
2
3 sed -i 's_https://www.udemy.com/__' udemy_courses_small.csv
4 sed -i 's_/__' udemy_courses_small.csv
5 sed -i -E 's/T[0-2][0-9]:[0-5][0-9]:[0-5][0-9]Z//' udemy_courses_small.csv
6
7 awk -F ',' -v OFS=',' '$9 = sprintf("%0.2f",$9)' udemy_courses_small.csv > udemy_courses_small_2.csv
8
```

Figure 1: Data cleaning by shell tools

7. Findings and discussion

The result is presented by each diagram on the dashboard. Grey square boxes on top of every chart signify the graph number as in Figure 2. Furthermore, the explanation of chart choice, filter, and color will also be given in the findings.

Graph 1: The first table in the dashboard gives general information about the total numbers of subscribers and reviews as well as a more detailed breakdown of free and paid courses. Around 3.5 million people (30.5%) attended the gratis classes while 8 million subscribers (69.5%) in the charged courses. However, only 571 thousand reviews were made, by far fewer than 11.6 million subscribers. Notably, there were only 301 non-paid versus 3,375 paid courses in Udemy, which occupied respectively 8.4% and 91.6% of the total courses. Although many users spent money on classes, data exhibit a substantial number of paid courses compared to free courses in Udemy.

Graph 2: Next, I present the count of courses in the four subjects: business finance, graphic design, music instruments, web development. The four subjects are divided into four distinctive colors, considering audiences with color blindness. Importantly, this color system for subjects is applied for other charts in the dashboard and our main filter for the whole dashboard. As can be seen from the graph, the course numbers for web development and business finance are almost similar, which are 1,200 and 1,196 respectively.

Graph 3: It is crucial showing the sum of subscribers and the average of subscribers side-by-side to compare the subject popularity. The y-axis is hidden in this chart because the subjects are grouped by color, as I mentioned above. On the left graph, the yellow squares present the number of reviews. Overall, with the highest subscribers and reviews, web development is the most popular subject while music instruments stay on the bottom with the least participants. An interesting observation is that the subject ranking swapped between business finance and graphic design when I shifted from the sum of subscribers to the average of subscribers. To be more specific, business finance ranks as the second winner with total subscribers with many courses but drops to third place in terms of average subscribers. The result explains more subscribers in total for business finance but, on average, fewer attendees per business finance course than graphic design. This could be considerable information for course creators to think about number of students might join their class with respect to the subject.

Graph 4: The scatter plot aims to examine if there is a relationship between the number of subscribers and content duration (in minutes). However, an analysis of the graph illustrates no clear association between the two variables. Hence, the content duration does not correlate to the popularity of the course.

Graph 5: The word cloud depicts the most popular course titles corresponding to the letter size through their subscribers. The range of subscriber numbers per course is used in this case to select more favorite courses. As it is demonstrated, "Learning HTML 5 Programming From Scratch" is the most frequent course. Moreover, most top courses are blue, which implies that they belong to the web development subject. Two courses, "Piano for all" and "Free Beginner Electric Guitar Lessons," are subject to musical instruments. By looking at the red title, only one business finance course is listed here with an absurd name: "Bitcoin or how I learned to stop worrying and love crypto."

Graph 6: The packed bubble diagram reveals the course pattern in a relational value with a cluster of circles. Each circle stands for one specific course. The circle color represents subjects; the subscriber number determines the circle size. The primary purpose of this plot is to get a general overview of the available courses in terms of their number, subject, and subscribers. It allows grasping the diversity of the course catalog in the blink of an eye. Moreover, the filter here denotes and lets us choose the number of subscribers per course. This filter only connects to this chart because the number of subscribers per course is not similar to the sum of subscribers from the other graphs. By hovering to the bubble, the audience can easily track the price and course titles. Many of the biggest circles are free courses.

Graph 7: The bar chart illustrates the average price, and the line graph demonstrates the number of subscribers by publication years from 2011 to 2016. The reason for excluding 2017 is that the dataset does not have the complete information for this whole year. There is no apparent similar trend between average price and quantity of attendees. The average price fluctuates amongst different years. On the other hand, the subscribers rose every year, particularly in 2015 but then fell slightly in 2016.

Graph 8: The last bar chart exhibits the number of subscribers in connection with course levels. Here all levels mean that these courses do not divide into any level that everyone can attend. The percentage of each subject in each level is also shown in this graph. Generally, all levels have the highest subscriber number, followed by beginner level, then the intermediate level, and lastly, the expert level.

8. Reflection and limitations

The project goals are to apply the theoretical foundation knowledge in the module into the practical project made by ourselves. To complete the mission, I proactively researched different data sources, chose a dataset, created meaningful research questions, cleaned data using various shell tools, established an interactive Tableau dashboard, and finally showed results based on revealed visualization. To sum up, there are subjects related to programming and development that might attract more users to an online platform like Udemy. Besides, I did not find any relationship between price or content duration to the number of subscribers. Additionally, Udemy provides way more paid classes than free courses. Nonetheless, free courses are crucial in attracting users to join Udemy platform since many courses with more subscribers are without fee. Finally, all levels or beginners have more participants than the intermediate and expert levels.

With many practices before finalizing the dashboard, I learned a lot from my mistakes and performances. However, shell tools are new, so it was time-consuming for me to practice and clean the dataset. Tableau is a user-friendly tool, so I tried to make our best intuitive dashboard. Data visualization requires much thought and analytical skills. As it is my first time working with Tableau, I have acquired significantly by practicing creating several dashboards. It is not only about Tableau but also data comprehension, storytelling, chart choices, logical presentation, suitable colors, helpful filters, and tooltips, data analysis and main messages.

Although I contribute the best of my work in this paper, the analysis remains some limitations. One of the limitations is that the dataset only contains data points from 2011 to the middle of 2017, while I would like to have the complete insights of Udemy courses till the current time. Secondly, Udemy currently offers around 13 subjects, but I can only work on four main subjects mentioned in the dataset (business finance, graphic design, musical instruments, and web design). Thirdly, the course popularity can be measured by other factors such as course quality, ratings, teacher reputation, but there is no available information on these variables in our dataset. Fourthly, our analysis concentrates on a descriptive study based on the actual data. The users can have the idea of course names, subjects, duration, levels, and the current pattern of subscribers over time; nevertheless, the predictive marks of subscriber number per course, per subject in the future is not the central point of analysis.

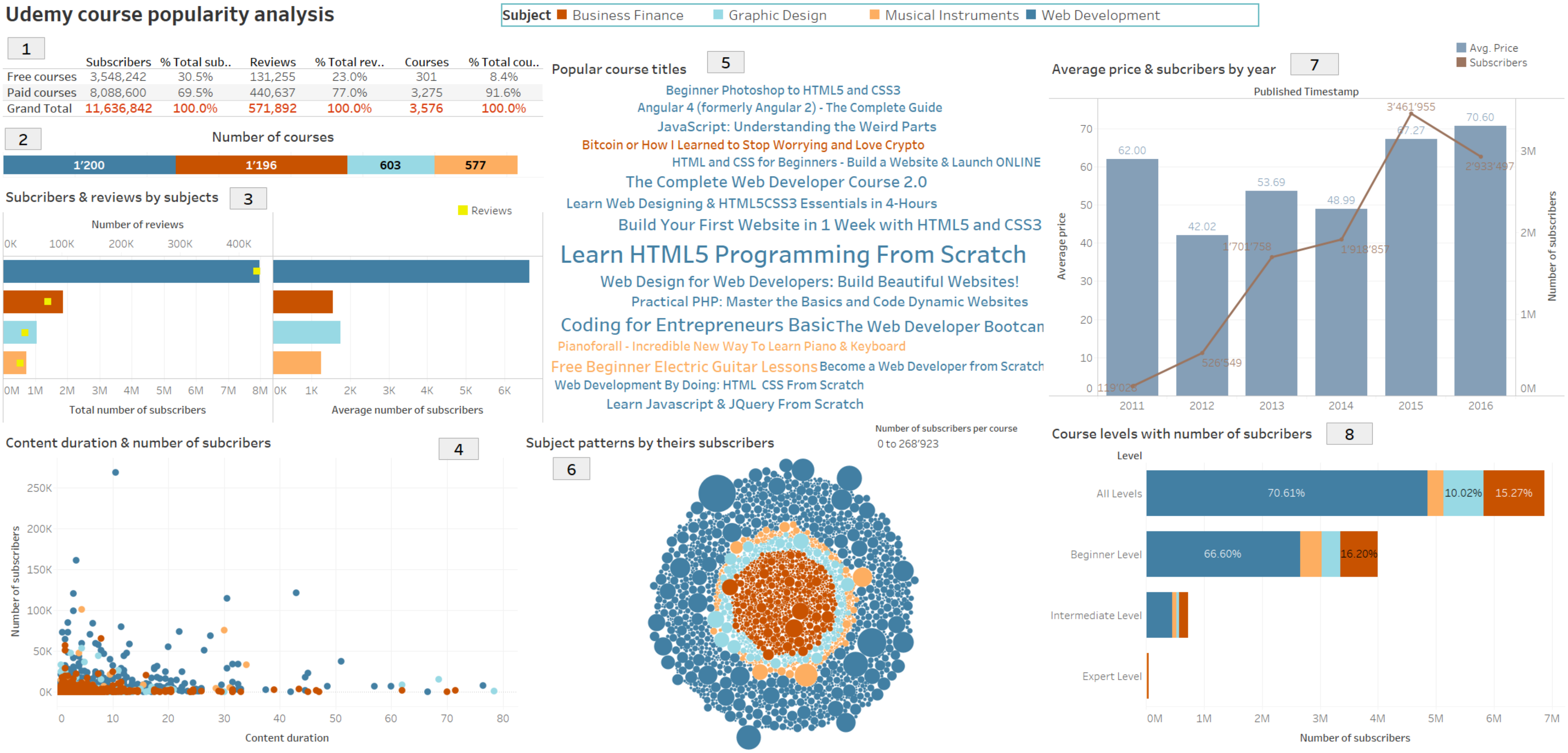
For future data analysis, it is better to have a more complete dataset that contains data points up to date with all subjects to gain better insight from data. It would be easier to make forecasts and predictions that the audiences get more interested in the analysis results.

Bibliography

Larxel. (2020, May 17). Udemy Courses. Retrieved April 16, 2021, from https://www.kaggle.com/andrewmvd/udemy-courses

Udemy (2021). Professional online course collection. Udemy for Business. Retrieved June 1, 2021, from https://business.udemy.com/course-collection/

Screenshot of dashboard



Udemy course popularity analysis

Subject 🟥 Business Finance 🟦 Graphic Design 🟧 Musical Instruments 🟦 Web Development

1

	Subscribers	% Total sub..	Reviews	% Total rev..	Courses	% Total cou..
Free courses	3,548,242	30.5%	131,255	23.0%	301	8.4%
Paid courses	8,088,600	69.5%	440,637	77.0%	3,275	91.6%
Grand Total	11,636,842	100.0%	571,892	100.0%	3,576	100.0%

2

Number of courses

Subcribers & reviews by subjects 3

Popular course titles 5

Beginner Photoshop to HTML5 and CSS3

Angular 4 (formerly Angular 2) - The Complete Guide

JavaScript: Understanding the Weird Parts

Bitcoin or How I Learned to Stop Worrying and Love Crypto

HTML and CSS for Beginners - Build a Website & Launch ONLINE

The Complete Web Developer Course 2.0

Learn Web Designing & HTML5CSS3 Essentials in 4-Hours

Build Your First Website in 1 Week with HTML5 and CSS3

Learn HTML5 Programming From Scratch

Web Design for Web Developers: Build Beautiful Websites!

Practical PHP: Master the Basics and Code Dynamic Websites

Coding for Entrepreneurs Basic The Web Developer Bootcan

Pianoforall - Incredible New Way To Learn Piano & Keyboard

Free Beginner Electric Guitar Lessons Become a Web Developer from Scratch

Web Development By Doing: HTML CSS From Scratch

Learn Javascript & JQuery From Scratch

Average price & subcribers by year 7

Content duration & number of subcribers 4

Subject patterns by theirs subscribers 6

Course levels with number of subcribers 8

Figure 2: Dashboard screenshot