UDEMY STUDENT BEHAVIOR ANALYSIS TO BUILD ONLINE COURSE

D4E65 - GROUP 1.1

Overview

- Story
- Dataset Overview
- Data Preprocessing
- Data Insight
- Conclusion

Story

– Our group wants to earn passive income from the online teaching platform Udemy by creating an attractive course.

– To build a successful course, we believe analyzing data from Udemy is extremely important. This analysis will help us answer how we should build such a course to attract many participants.

Dataset Overview

udemy_courses dataset from GitHub by MainakRepositor

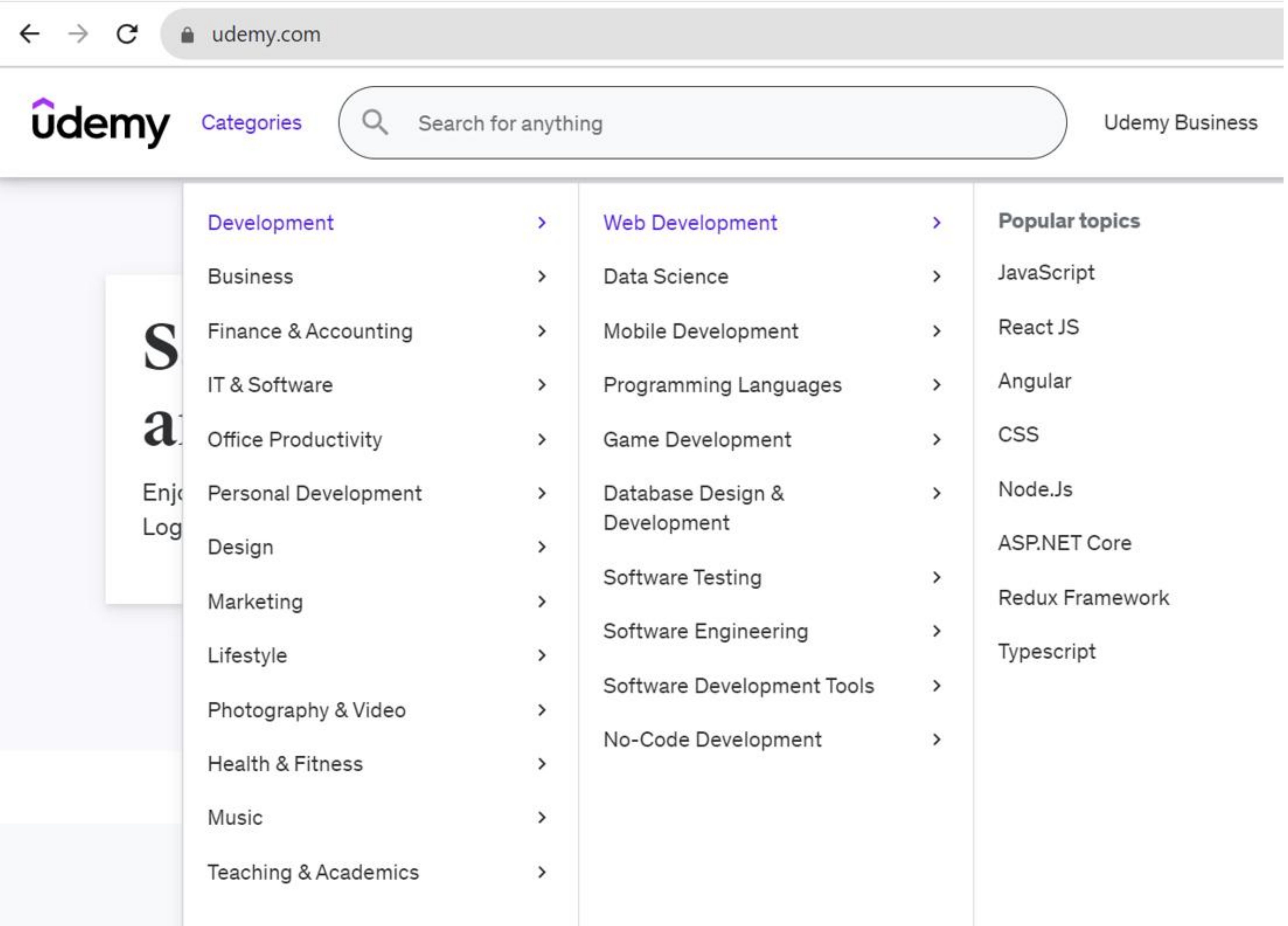
3678 rows x 12 columns

Link

No.	Column name	Description
1	course_id	Course ID
2	course_title	Course title
3	url	Course url
4	is_paid	Paid or Free course (True/False)
5	price	Course price (USD unit)
6	num_subscribers	Number of subscribers per course
7	num_reviews	Number of reviews per course
8	num_lectures	Number of lectures per course
9	level	Beginner/Intermediate/Expert/All Levels
10	content_duration	Total hours of the course
11	published_timestamp	Course publish date
12	subject	Business Finance/Graphic Design/Musical Instruments/Web Development

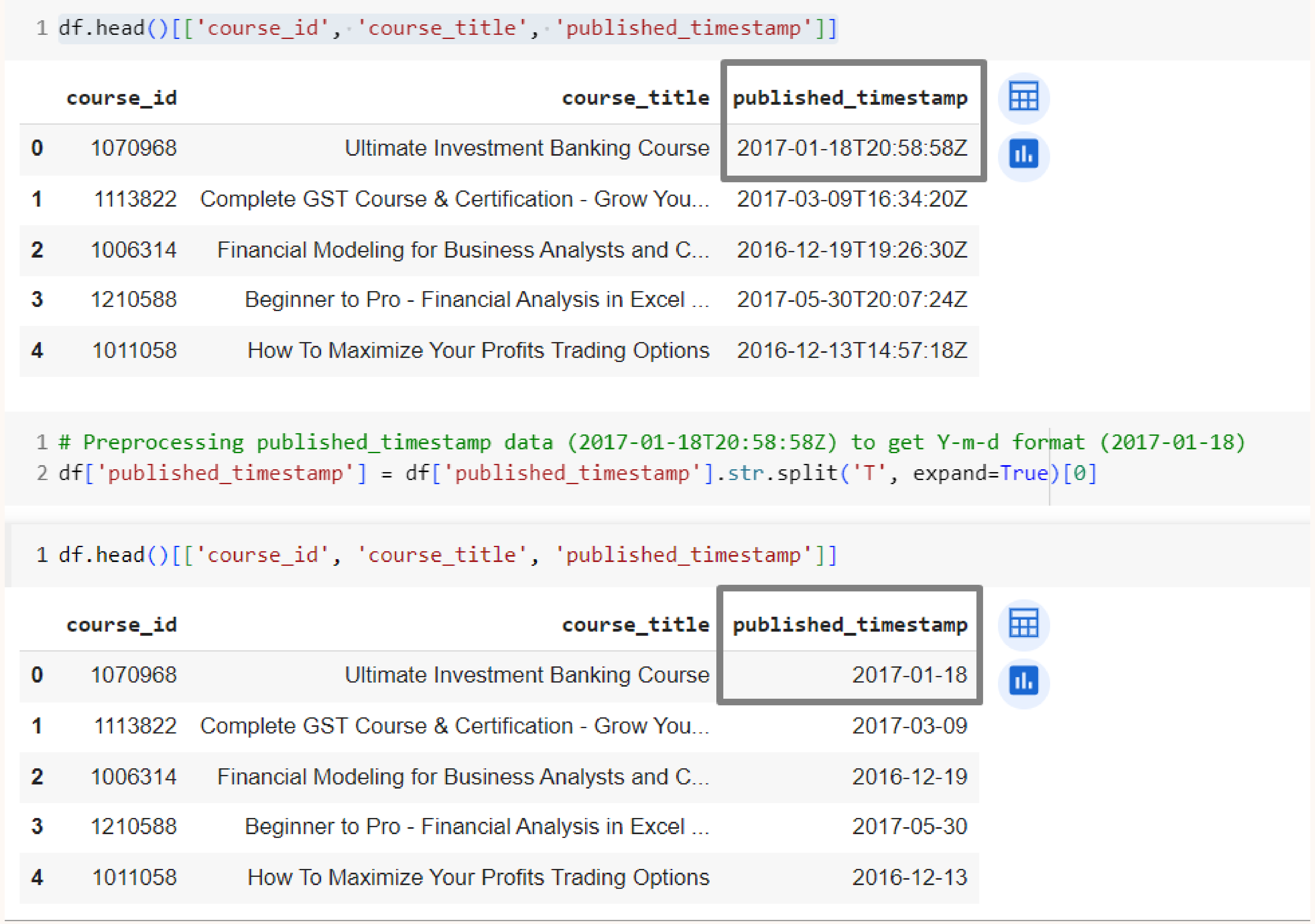
Key of Dataset

- **content_duration**
- **level**
- **price**
- **num_lectures**
- **subject**
- **topic**



Data Preprocessing

Preprocessing published_timestamp data (2017-01-18T20:58:58Z) to get Y-m-d format (2017-01-18)



```
df.head()[['course_id', 'course_title', 'published_timestamp']]
```

	course_id	course_title	published_timestamp
0	1070968	Ultimate Investment Banking Course	2017-01-18T20:58:58Z
1	1113822	Complete GST Course & Certification - Grow You...	2017-03-09T16:34:20Z
2	1006314	Financial Modeling for Business Analysts and C...	2016-12-19T19:26:30Z
3	1210588	Beginner to Pro - Financial Analysis in Excel ...	2017-05-30T20:07:24Z
4	1011058	How To Maximize Your Profits Trading Options	2016-12-13T14:57:18Z

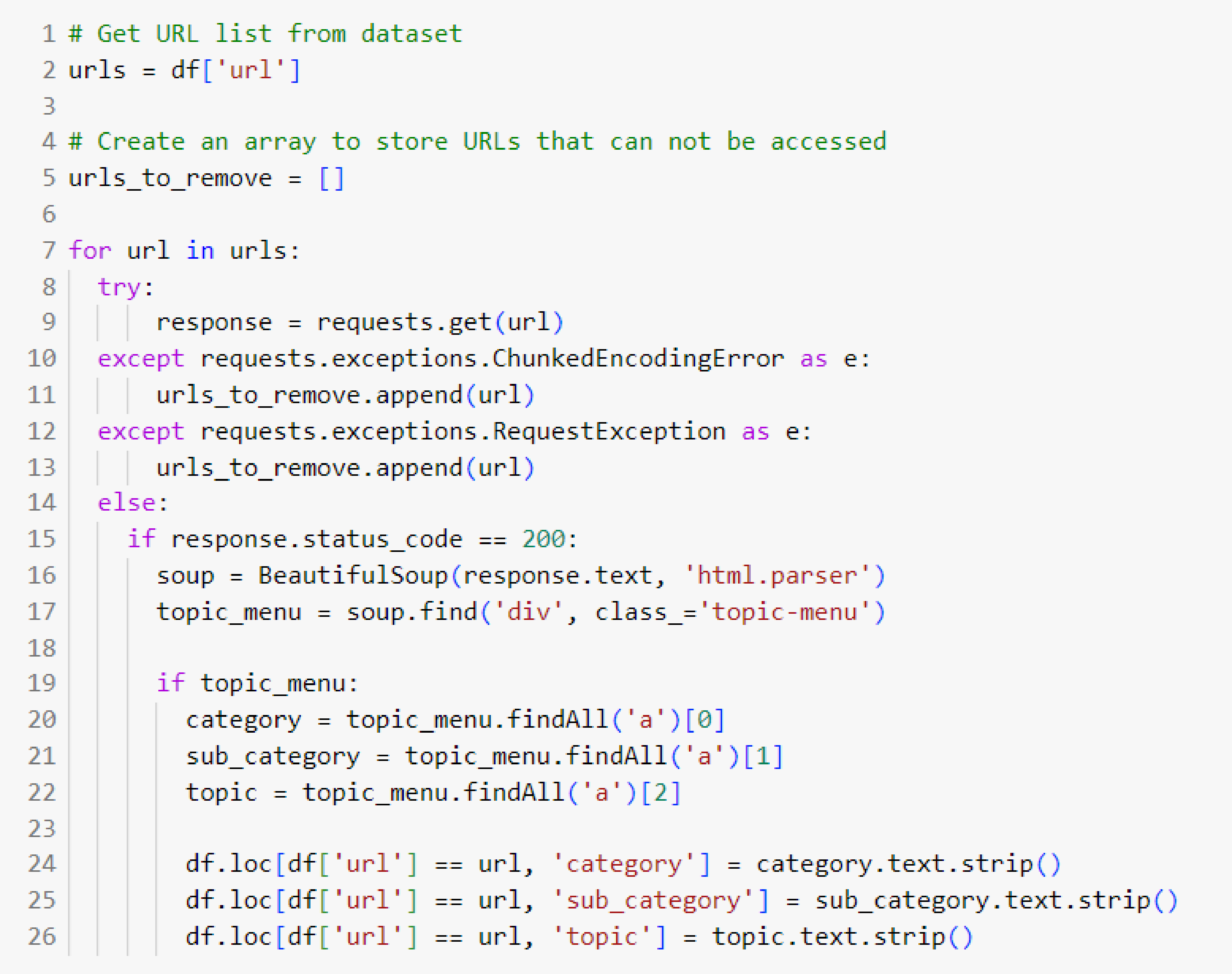
```
# Preprocessing published_timestamp data (2017-01-18T20:58:58Z) to get Y-m-d format (2017-01-18)
df['published_timestamp'] = df['published_timestamp'].str.split('T', expand=True)[0]
```

```
df.head()[['course_id', 'course_title', 'published_timestamp']]
```

	course_id	course_title	published_timestamp
0	1070968	Ultimate Investment Banking Course	2017-01-18
1	1113822	Complete GST Course & Certification - Grow You...	2017-03-09
2	1006314	Financial Modeling for Business Analysts and C...	2016-12-19
3	1210588	Beginner to Pro - Financial Analysis in Excel ...	2017-05-30
4	1011058	How To Maximize Your Profits Trading Options	2016-12-13

Data Preprocessing

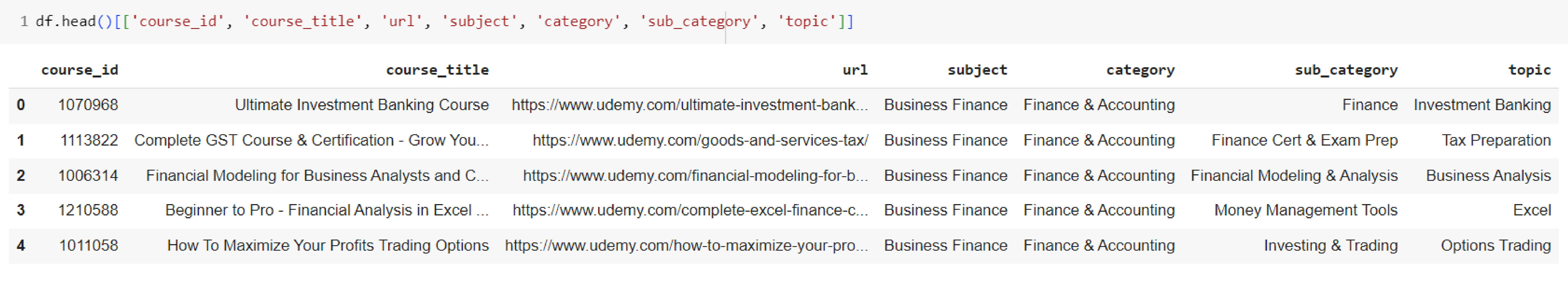
Crawling data base on urls from dataset



```python
# Get URL list from dataset
urls = df['url']

# Create an array to store URLs that can not be accessed
urls_to_remove = []

for url in urls:
    try:
        response = requests.get(url)
    except requests.exceptions.ChunkedEncodingError as e:
        urls_to_remove.append(url)
    except requests.exceptions.RequestException as e:
        urls_to_remove.append(url)
    else:
        if response.status_code == 200:
            soup = BeautifulSoup(response.text, 'html.parser')
            topic_menu = soup.find('div', class_='topic-menu')

            if topic_menu:
                category = topic_menu.findAll('a')[0]
                sub_category = topic_menu.findAll('a')[1]
                topic = topic_menu.findAll('a')[2]

                df.loc[df['url'] == url, 'category'] = category.text.strip()
                df.loc[df['url'] == url, 'sub_category'] = sub_category.text.strip()
                df.loc[df['url'] == url, 'topic'] = topic.text.strip()
```

Data Preprocessing

Result: A new dataset with 3678 rows x 15 columns



```python
df.head()[['course_id', 'course_title', 'url', 'subject', 'category', 'sub_category', 'topic']]
```

	course_id	course_title	url	subject	category	sub_category	topic
0	1070968	Ultimate Investment Banking Course	https://www.udemy.com/ultimate-investment-bank...	Business Finance	Finance & Accounting	Finance	Investment Banking
1	1113822	Complete GST Course & Certification - Grow You...	https://www.udemy.com/goods-and-services-tax/	Business Finance	Finance & Accounting	Finance Cert & Exam Prep	Tax Preparation
2	1006314	Financial Modeling for Business Analysts and C...	https://www.udemy.com/financial-modeling-for-b...	Business Finance	Finance & Accounting	Financial Modeling & Analysis	Business Analysis
3	1210588	Beginner to Pro - Financial Analysis in Excel ...	https://www.udemy.com/complete-excel-finance-c...	Business Finance	Finance & Accounting	Money Management Tools	Excel
4	1011058	How To Maximize Your Profits Trading Options	https://www.udemy.com/how-to-maximize-your-pro...	Business Finance	Finance & Accounting	Investing & Trading	Options Trading

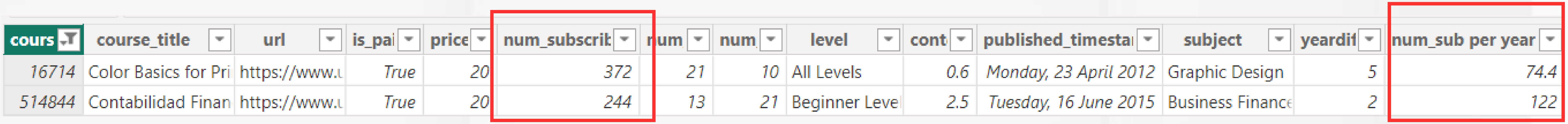
New columns:

- category
- sub_category
- topic

- ***Dataset from 2011 to 2017:***

3678	12m	443k	310
Total Number of Course	Total Number of Subscribes	Total Number of Reviews	Total Free Courses

Due to the fact that there are courses created from a long time ago, such as 2011, 2012, etc. But if analyzed in such a general way, it seems that the data will not be correct . Therefore, I will consider further analyzing the data by year.



cours	course_title	url	is_pai	price	num_subscrit	num	num	level	cont	published_timestai	subject	yeardif	num_sub per year
16714	Color Basics for Pri	https://www.u	True	20	372	21	10	All Levels	0.6	Monday, 23 April 2012	Graphic Design	5	74.4
514844	Contabilidad Finan	https://www.u	True	20	244	13	21	Beginner Leve	2.5	Tuesday, 16 June 2015	Business Finance	2	122

DATA ANALYSIS ACCORDING TO EACH YEAR'S AVERAGE

658

Total Number of Course per year

2.42m

Total Number of Subscribes per year

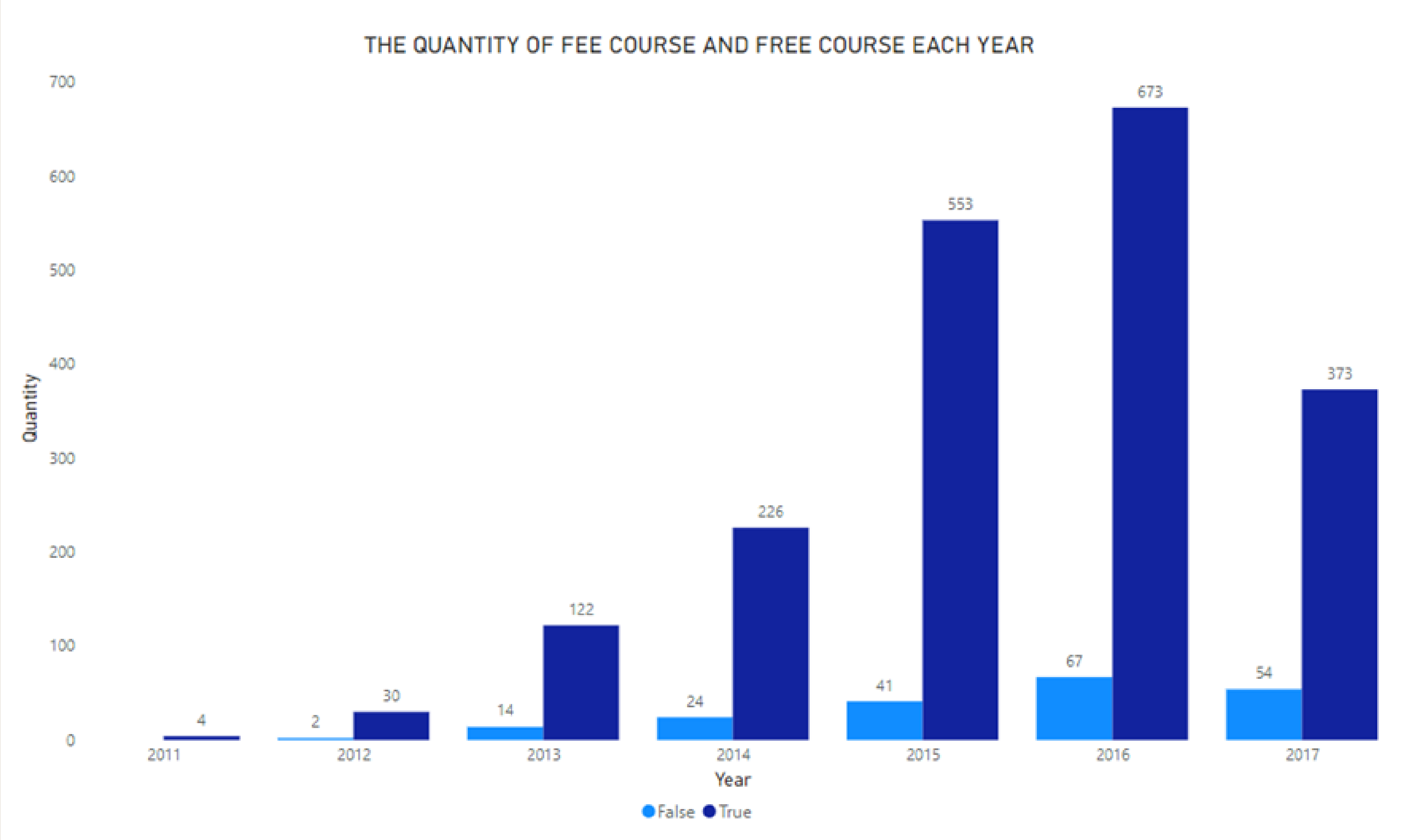
122k

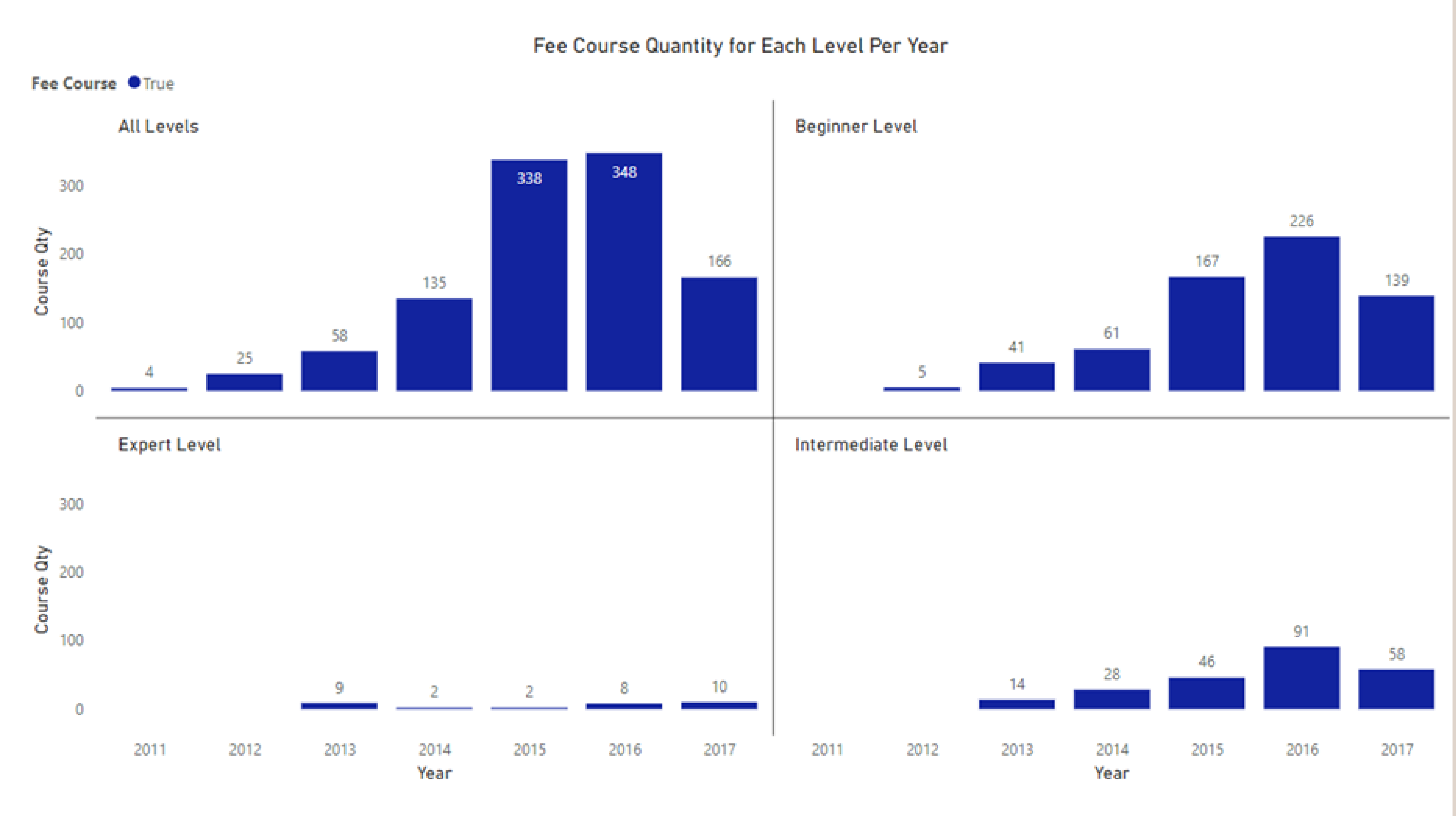
Total Number of Reviewes per year

310

Total Free Course per year

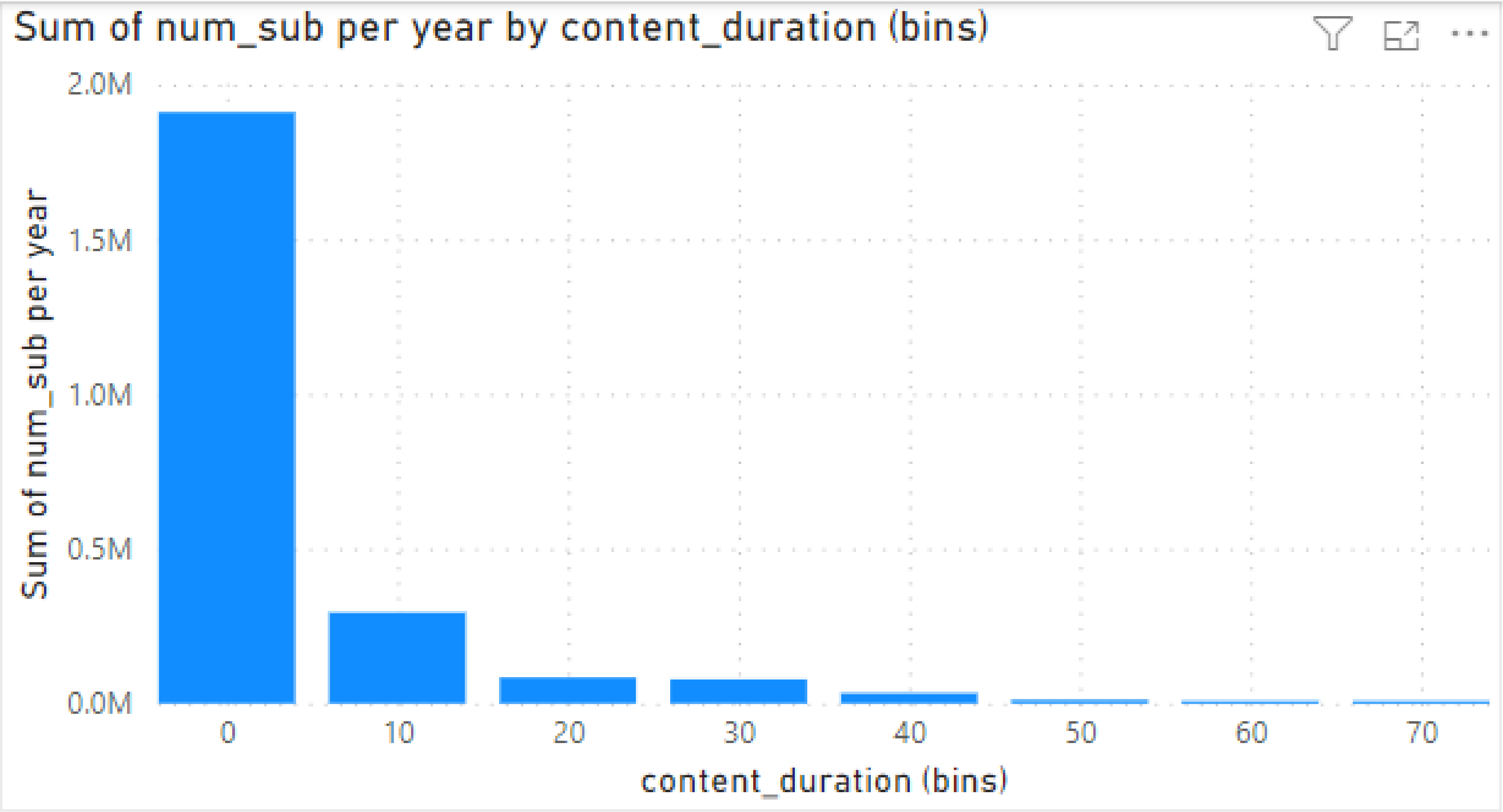
Before building a online course, we need to know what exactly behavior of learner. By sorting which course are fee or free and the level is the most chosen by learner.





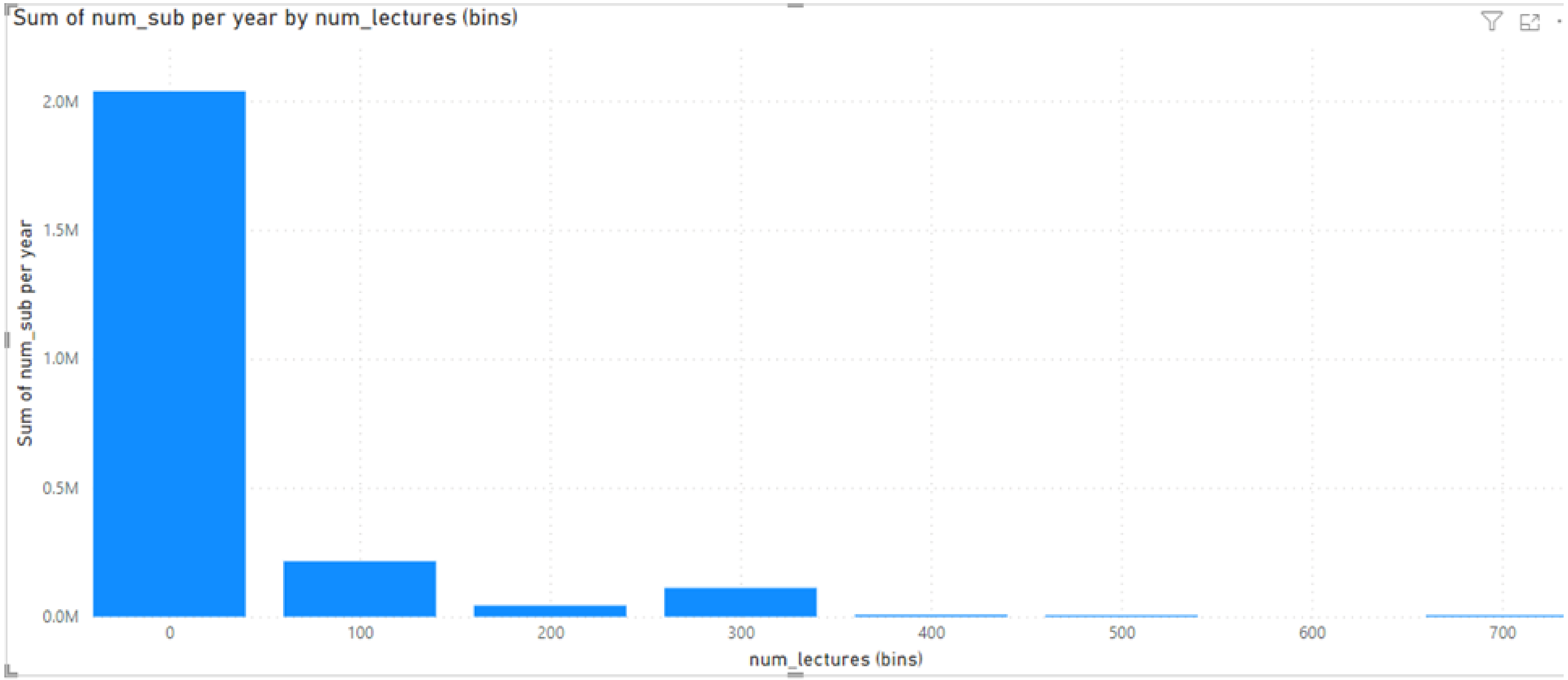
The above charts are showed that the learner choose FEE COURSE more than FREE COURSE and beside that the level attracts to them are ALL LEVEL and BEGINNER

The course content timing will also take a main feature for building a online course. Look at detail on the below chart, you see...



The courses duration content less than 10 HOURS have got attraction from the learner

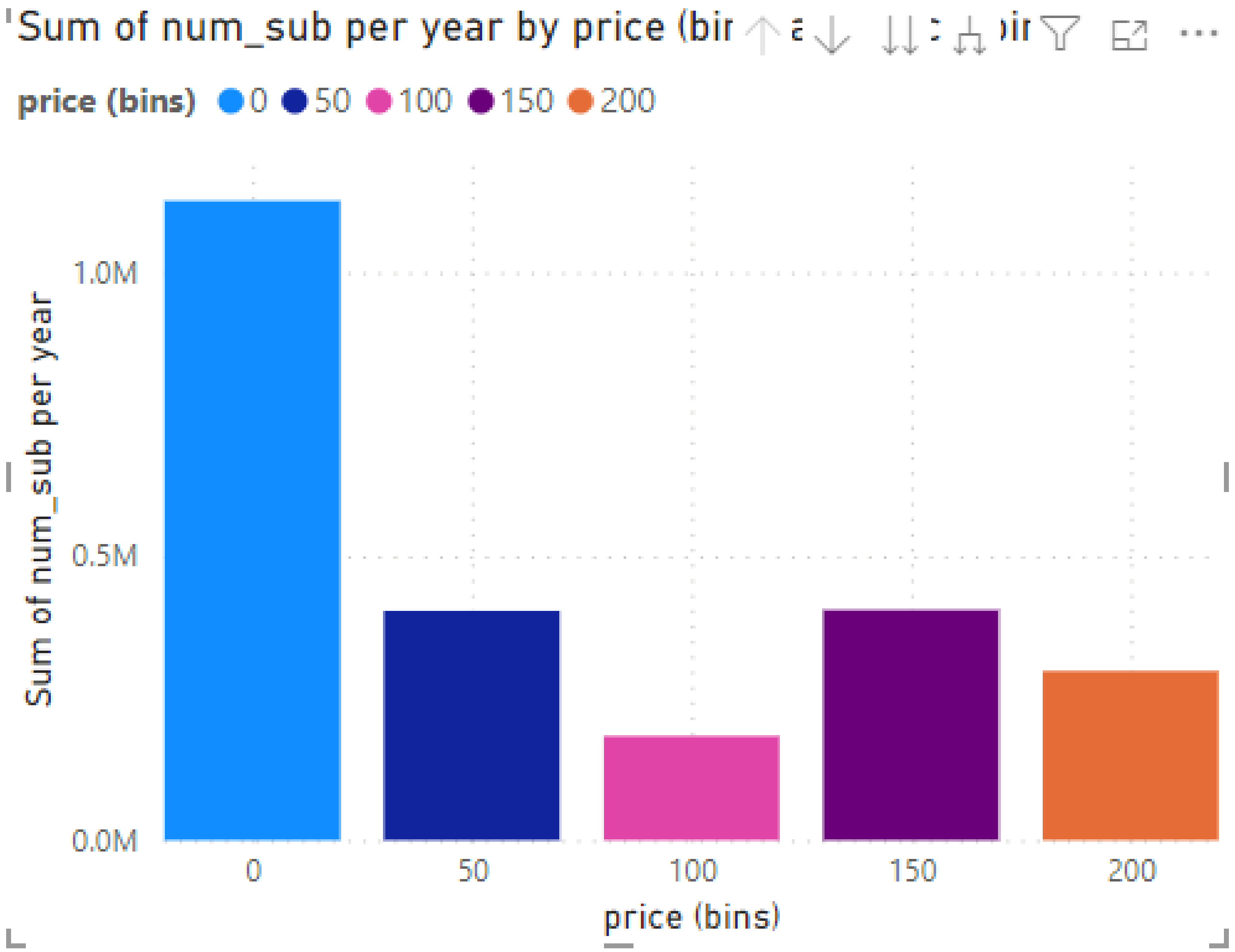
Following the course duration content, our group move to analyze how many lecture are suitable for learner. And we realize that,...

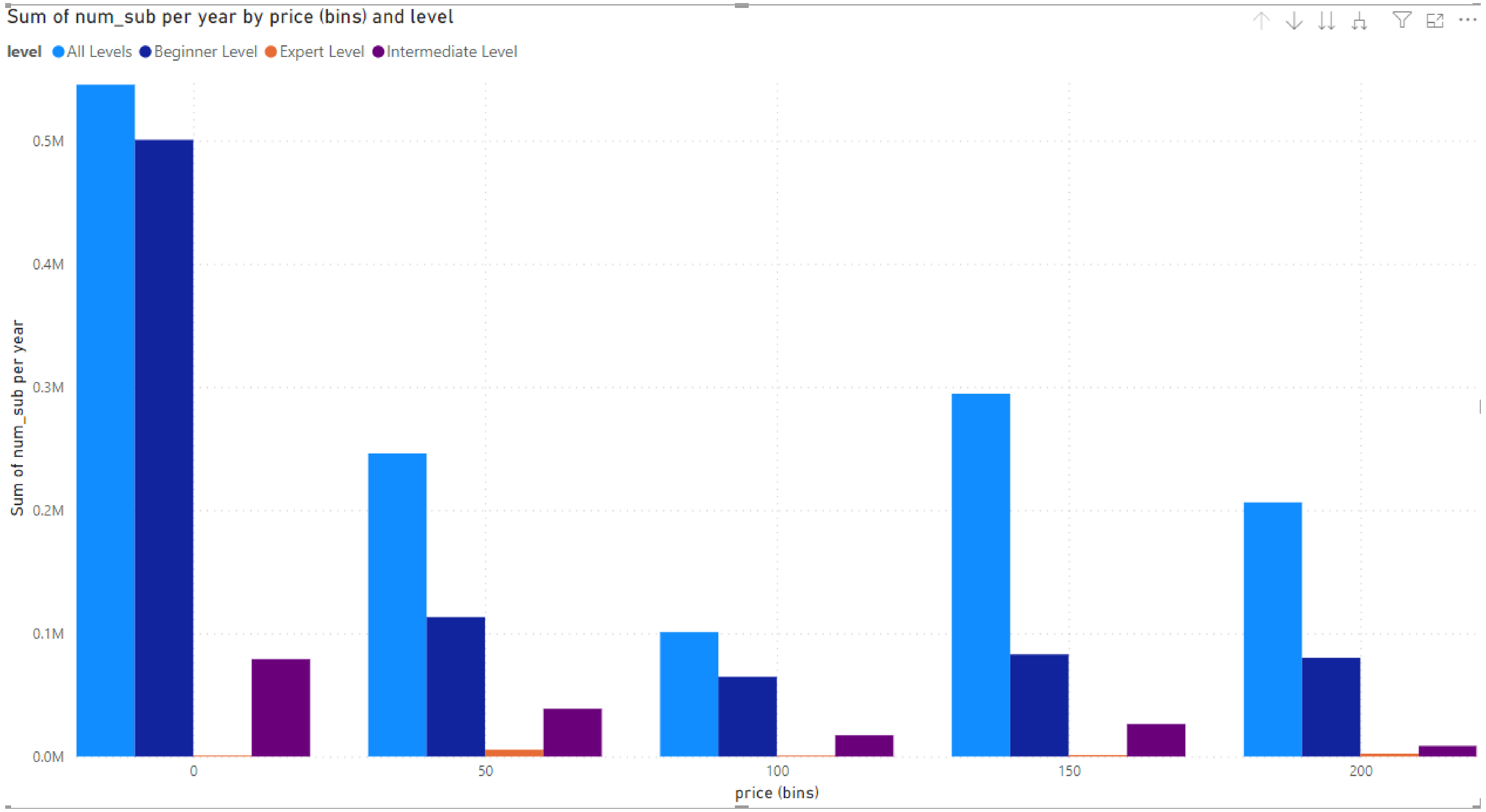


NOT OVER 100 LECTURES **are the trend for learner when they are attending the online courses**

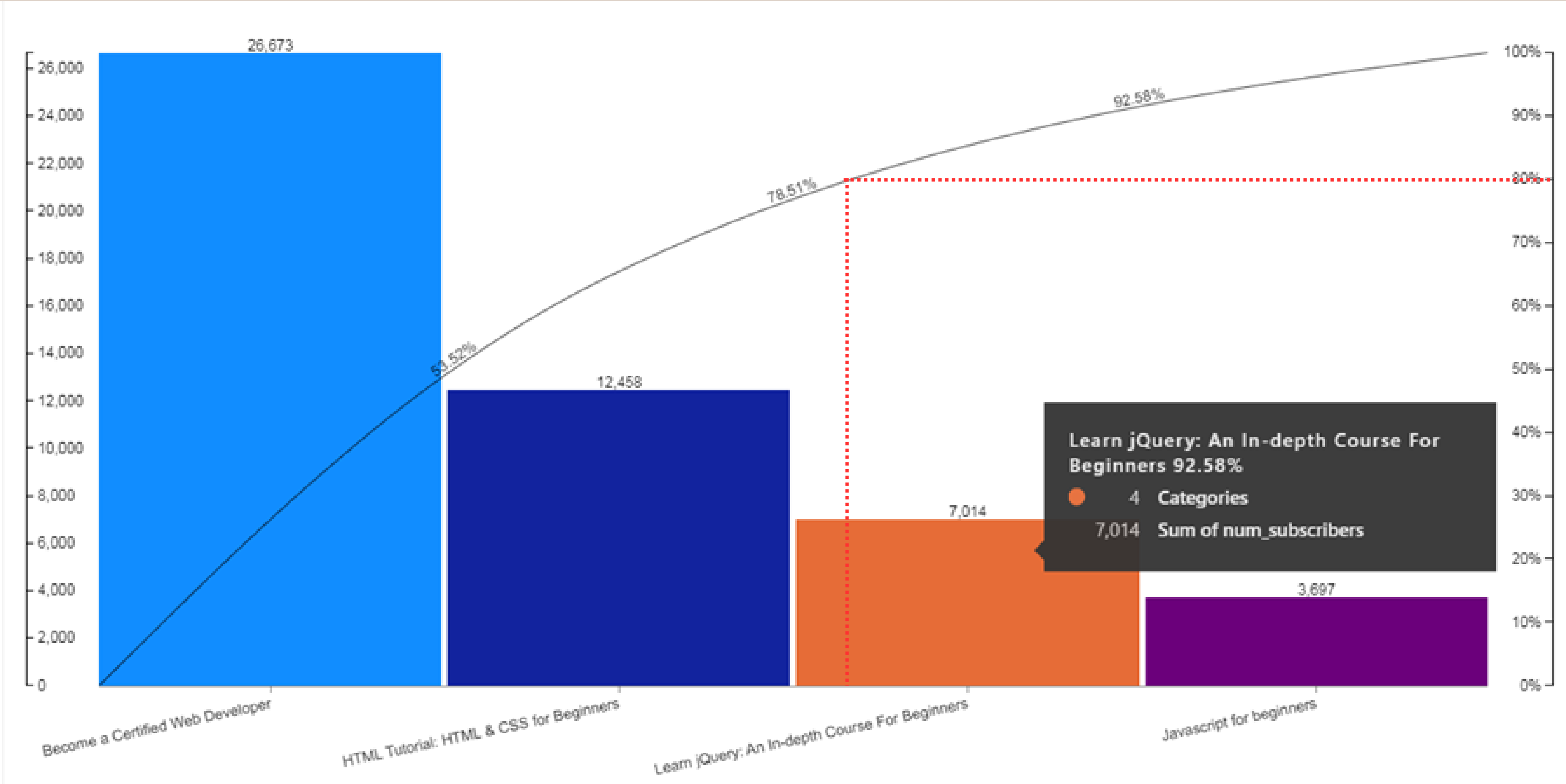
Insight 1: The price range we should use for a course is range 1, 2, and 4.

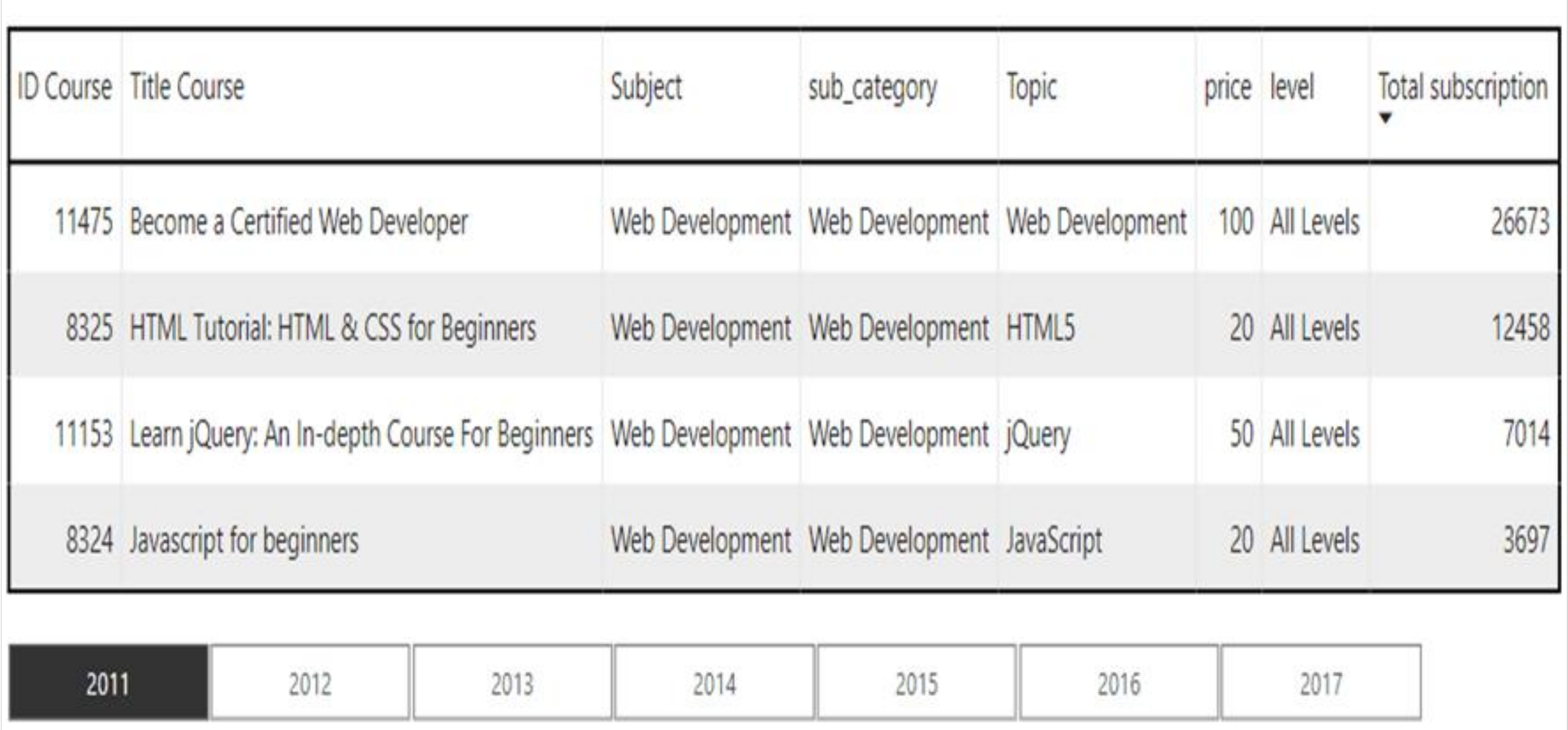
Insight 2: All Levels courses can be priced from 150 to 199 USD, and it is not necessary to lower the price to compete.





After analyzing behavior and finance of the learner, our group can create a structure basic online course. Now we will validate and select what subject and topic is the most suitable for online course investment.

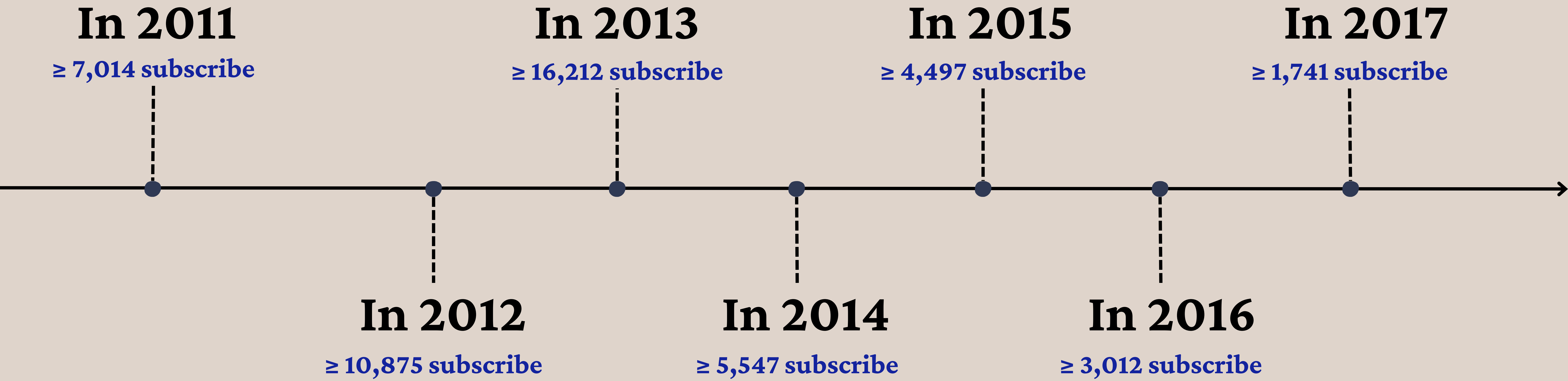




ID Course	Title Course	Subject	sub_category	Topic	price	level	Total subscription ▼
11475	Become a Certified Web Developer	Web Development	Web Development	Web Development	100	All Levels	26673
8325	HTML Tutorial: HTML & CSS for Beginners	Web Development	Web Development	HTML5	20	All Levels	12458
11153	Learn jQuery: An In-depth Course For Beginners	Web Development	Web Development	jQuery	50	All Levels	7014
8324	Javascript for beginners	Web Development	Web Development	JavaScript	20	All Levels	3697

Applying the 80:20 RULE methodology (it is also called PARETO CHART) to select which course have high subscribe per year. The above charts illustrate that we will choose what subject have the amount of subscribe more than 7,014 in 2011.

Similarly in 2011, our group continuously apply 80:20 rule for another year. And get the result that...



In 2011

≥ 7,014 subscribe

In 2012

≥ 10,875 subscribe

In 2013

≥ 16,212 subscribe

In 2014

≥ 5,547 subscribe

In 2015

≥ 4,497 subscribe

In 2016

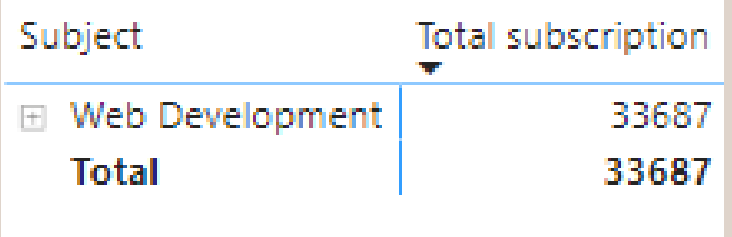
≥ 3,012 subscribe

In 2017

≥ 1,741 subscribe

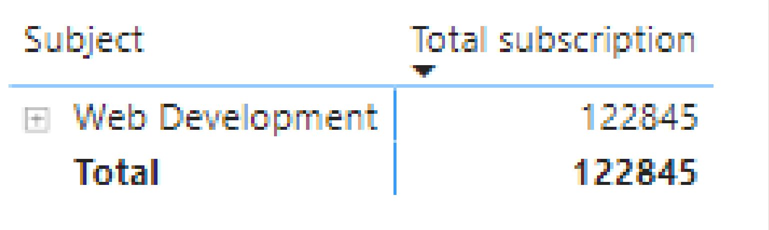
When getting top courses were subscribed by the learner. Move to next step is sorting what subject is more popular

In 2011



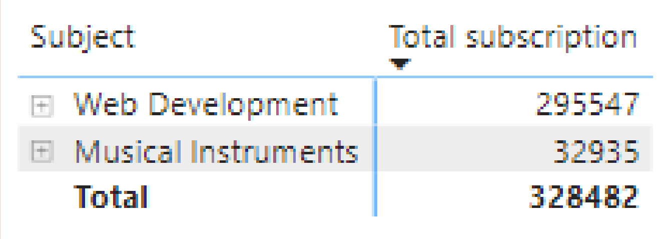
Subject	Total subscription
Web Development	33687
Total	**33687**

In 2012



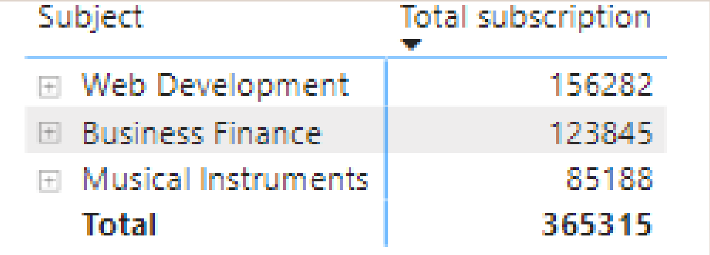
Subject	Total subscription
Web Development	122845
Total	**122845**

In 2013



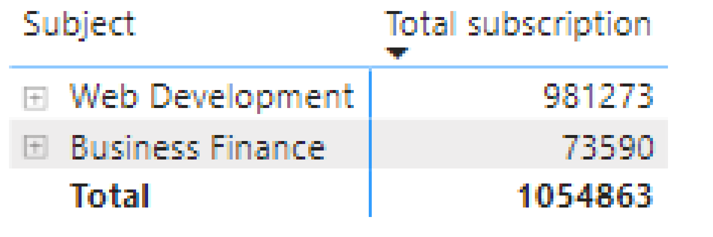
Subject	Total subscription
Web Development	295547
Musical Instruments	32935
Total	**328482**

In 2014



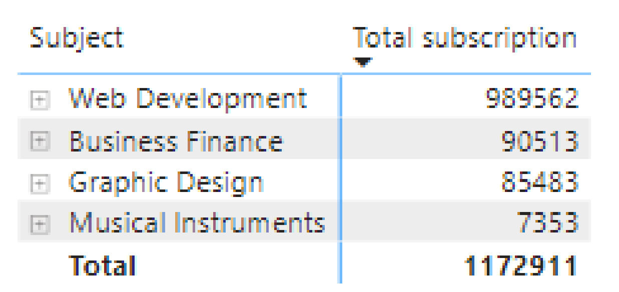
Subject	Total subscription
Web Development	156282
Business Finance	123845
Musical Instruments	85188
Total	**365315**

In 2015



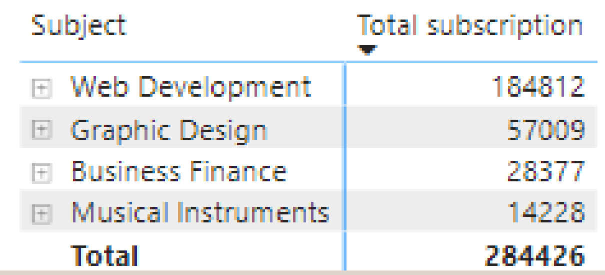
Subject	Total subscription
Web Development	981273
Business Finance	73590
Total	**1054863**

In 2016



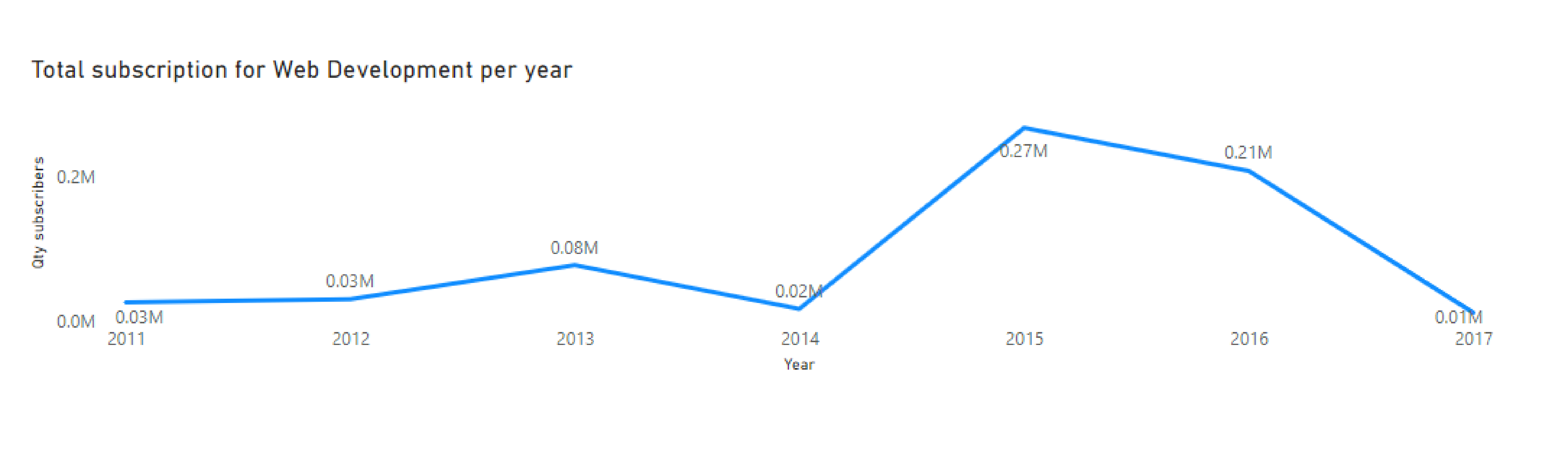
Subject	Total subscription
Web Development	989562
Business Finance	90513
Graphic Design	85483
Musical Instruments	7353
Total	**1172911**

In 2017

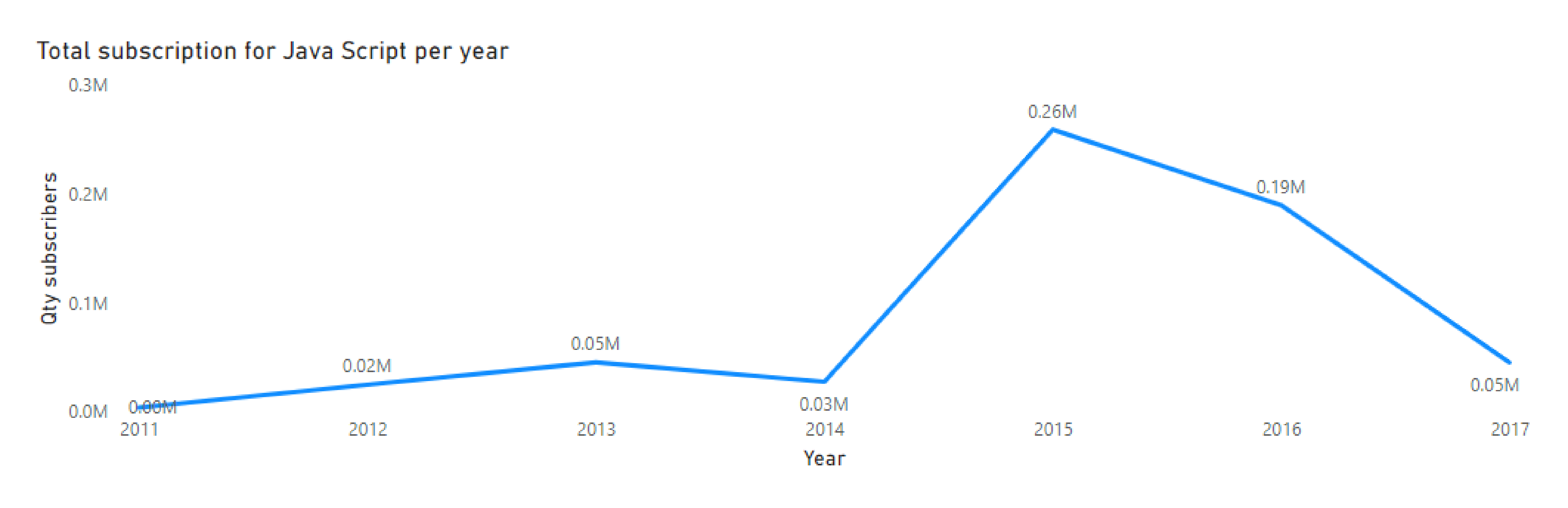


Subject	Total subscription
Web Development	184812
Graphic Design	57009
Business Finance	28377
Musical Instruments	14228
Total	**284426**

Choosing TOPIC which more subscribed


Total subscription for Web Development per year

✗


Total subscription for Java Script per year

✓

Conclusion

- **content_duration**

  Less than 10 hours

- **num_lectures**

  Not over 100 lectures

- **level**

  All level

- **subject**

  Web development

- **price**

  From 150 to 199 USD

- **topic**

  Java Script

THANK YOU