UDEMY: EVALUATION OF THE GROWTH OF COURSES

George Mason University

AIT-580-001 | Prof. Dr. Harry Foxwell

Anil Kumar Mureboina

George Mason University

Fairfax, Virginia

amureboi@gmu.edu

ABSTRACT

There are several MOOC (Massive Open Online Courses) platforms available on the internet, including Udemy, Coursera, LinkedIn Learning, edX, and others. I can assure you that the number of users on those three sites has risen dramatically in recent years. I must agree that the availability of those MOOC sites is beneficial to students. I selected Udemy for this project since it is my favorite online learning site. The primary goal of this paper is to investigate the relevance and possible advantages of Udemy. This dataset covers information on Udemy from 2011 to 2017. This dataset includes information about several courses, such as course price, duration, rating, number of lectures, etc. The dataset includes information on the proficiency level of each category and publication dates. Analyzing this dataset provides students with a better understanding of the importance of online learning platforms and how to focus on certain courses depending on their preferences. Students can choose a course depending on the pricing, duration, and popularity of the course.

Keywords - Udemy, Category, Level, Price, Subscribers, Date, Python, R, SQL

I. INTRODUCTION

UDEMY:

Udemy is an online learning platform that has over 120,000 individual online courses. It provides courses delivered by instructors via video learning modules. Development, business, finance, and accounting, IT & software, office productivity, personal development, design, marketing, lifestyle, photography & video, health & fitness, music, and teaching & academics are among the platform's offerings.

Every year, millions of people choose online platform courses to get qualitative and practical information. The goal of this research is to show how online platforms work to deliver competitive education. I'd want to talk about the dataset's relevance and possible advantages. This data must be studied since it will enable students to focus on certain courses depending on their interests and talents. Students can choose a course depending on its cost and duration. Furthermore, by studying this data, the popularity of the course for the relevant disciplines can be identified, which assists in presenting the student with a long-term perspective on subscribing to a course. The rating tool allows us to pick courses intelligently and filter them based on their ratings, subscribers, and skill level. This dataset includes all the NOIR data types.

II. OBJECTIVE

The purpose of this article is to investigate the following research questions using data from Udemy:

1) What is the average price of the courses for all the categories?
2) What are the most expensive courses among all the categories?
3) What is the average duration of the course based on the proficiency level?
4) Which course has the most subscribers in each category?
5) Which course has the most reviews in each category?
6) What are the most popular courses and when they are published?
7) What is the number of proficiency levels in each course category?

III. LITERATURE REVIEW

ARTICLE 1:

ONLINE PLATFORM COURSES BETWEEN EDUCATION AND BUSINESS

Millions of people choose online platform courses to gain qualitative and practical knowledge every year. The purpose of this study is to depict the effort made by online platforms to provide competitive education. During 2020, they monitored the activity on the Coursera and Udemy platforms and took online courses in the Business category. They checked the offer of new courses on platforms monthly and tracked their popularity to get a sense of how well the platforms adapt to people's needs. Many people believe that open educational resources will increase access, lower costs, and improve educational quality. While Coursera has collaborated with prestigious universities to provide high-quality free courses over the years, Udemy relies on business professionals to provide its courses. Business, human resources, and project management courses range in price from 11,99 euros to 199,99 euros. They took the average number of courses by category from the Coursera portal and displayed the difference between individual courses. In addition, the average number of enrolled students increased across categories from 2019 to 2021. They also took a list of courses from Udemy in several categories, including Business, Personal development, Information technology, social sciences, Physical science and engineering, Math and logic and Business, Data science and health and Computer science, Art and humanities, Language learning, web development, Design, and Digital marketing and Photoshop. They displayed the number of courses available in each category as well as the average number of hours and cost of those courses. They created a graph depicting the number of participants in the most popular Udemy topics in 2020. The Udemy platform was founded on the combination of an online business and the need for flexible education at a low cost. The number of students enrolled increased significantly in 2020, and the courses have been continuously improved.[10]

ARTICLE 2:

DATA COURSE ANALYSIS AND VISUALIZATION AT UDEMY

Researchers obtained data from Udemy via Web Scraping. They went through seven steps to collect data and answers to their research questions. For data analysis, researchers use the

python programming language (Google collaboration) and blue for data visualization. They collected information such as the course category and count, as well as the number of subscribers for the courses. They also collected information on the cost of the courses as well as their ratings. With this data, they came to several conclusions about Udemy's capabilities and how it will add high-quality courses. Udemy can expand courses that fall into a financially attractive category. They added a scatter plot that shows the number of subscribers based on course category using the categories, their count, and subscribers data.[3]

ARTICLE 3:

LEGAL EDUCATION THROUGH MOOCS: A STUDY OF SELECT INTERNATIONAL ONLINE PLATFORMS

EdX, Coursera, Udemy, LinkedIn Learning, and SWAYAM are the five platforms that offer courses in the field of law. Udemy is the platform from which the majority of the courses are available. Of the 630 MOOCs listed in Law, 329 were offered through Udemy. It has also been discovered that most courses are offered for less than or equal to 48 hours. They provided information on the chosen MOOC platforms. From a total of 630 courses, they created a pie chart that reveals that 79% are paid, 14% are paid certification courses, and 7% are free courses. The bar graph shows that 346 (55%) of the courses are offered in English, with 86 (14%) being offered in Portuguese. Only two courses are available in Indonesian and Italian, with five courses being given in Chinese. Three courses in Arabic are offered after four each in Japanese and Polish languages. According to the analysis, many courses are available for less than or equal to 48 hours on the five selected platforms; most of these short-duration courses are also available on Udemy and LinkedIn Learning.[2]

IV. MATERIALS AND METHODS

A. Data Collection

The dataset utilized for the study was obtained from GitHub. This dataset was collected as a CSV file, making it easy to import into multiple software. This dataset contains information from Udemy from 2011 to 2017. This dataset contains information on the course title, course fee, duration, rating, number of lectures, course competence level of each category, and publishing dates. With a vast dataset with numerous variables, our study was able to obtain a wealth of information about the benefits of Udemy.[1]

B. Data Cleaning

Data cleaning plays a major role in any project. There is no value for the analysis without cleaning the data. Removal of missing data and irrelevant data is extremely important. The data of this project has been cleaned by using R and Python. Unnecessary data and missing values are cleaned using R and the string manipulation is done through Python to check the relationship between the variables. This dataset contains the id, title, price, content duration,

category, level, subscribers, reviews, and lectures of the course. The following figure is the data that has been loaded in SQL after cleaning.[4][5][8]

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
1070968	Ultimate Investment Banking Course	TRUE	200	2147	23	51	All Levels	1.5	2017-01-18 20:58:58	Business Finance
1113822	Complete GST Course & Certification - Grow Your CA Practice	TRUE	75	2792	923	274	All Levels	39	2017-03-09 16:34:20	Business Finance
1006314	Financial Modeling for Business Analysts and Consultants	TRUE	45	2174	74	51	Intermediate Level	2.5	2016-12-19 19:26:30	Business Finance
1210588	Beginner to Pro - Financial Analysis in Excel 2017	TRUE	95	2451	11	36	All Levels	3	2017-05-30 20:07:24	Business Finance
1011058	How To Maximize Your Profits Trading Options	TRUE	200	1276	45	26	Intermediate Level	2	2016-12-13 14:57:18	Business Finance
192870	Trading Penny Stocks: A Guide for All Levels In 2017	TRUE	150	9221	138	25	All Levels	3	2014-05-02 15:13:30	Business Finance
739964	Investing And Trading For Beginners: Mastering Price Charts	TRUE	65	1540	178	26	Beginner Level	1	2016-02-21 18:23:12	Business Finance
403100	Trading Stock Chart Patterns For Immediate, Explosive Gains	TRUE	95	2917	148	23	All Levels	2.5	2015-01-30 22:13:03	Business Finance
476268	Options Trading 3 : Advanced Stock Profit and Success Method	TRUE	195	5172	34	38	Expert Level	2.5	2015-05-28 00:14:03	Business Finance
1167710	The Only Investment Strategy You Need For Your Retirement	TRUE	200	827	14	15	All Levels	1	2017-04-18 18:13:32	Business Finance
592338	Forex Trading Secrets of the Pros With Amazon's AWS	TRUE	200	4284	93	76	All Levels	5	2015-09-11 16:47:02	Business Finance
975046	Trading Options With Money Flow	TRUE	200	1380	42	17	All Levels	1	2016-10-18 22:52:31	Business Finance
742602	Financial Management Risk and Return For Securities	TRUE	30	3607	21	19	All Levels	1.5	2016-02-03 18:04:01	Business Finance
794151	Forex Trading Course: Work Smarter Not Harder Proven Results	TRUE	195	4061	52	16	All Levels	2	2016-03-16 15:40:19	Business Finance
1196544	Python Algo Trading: Sentiment Trading with News	TRUE	200	294	19	42	All Levels	7	2017-04-28 16:41:44	Business Finance
504036	Short Selling: Learn To Sell Stocks Before The Fall	TRUE	75	2276	106	19	Intermediate Level	1.5	2015-06-22 21:18:35	Business Finance
719698	Basic Technical Analysis: Learn the structure of the market	TRUE	20	4919	79	16	Beginner Level	1.5	2016-01-08 17:21:26	Business Finance
564966	The Complete Chart Pattern Trading Course: A Proven Approach	TRUE	200	2666	115	52	All Levels	4	2015-08-10 21:07:35	Business Finance
606928	7 Deadly Mistakes of Investing that Will Slash Your Profits!	TRUE	50	5354	24	23	All Levels	1.5	2015-09-21 18:10:34	Business Finance
58977	Financial Statements Made Easy	TRUE	95	8095	249	12	Beginner Level	0.58333333333...	2013-06-09 00:21:26	Business Finance
1242604	Winning Forex Trading with Live Forex Trading Examples	TRUE	200	809	3	25	All Levels	2	2017-06-06 02:54:04	Business Finance
798740	Forex Traders - Create Your First Trading Robot. No Coding!	TRUE	200	2295	84	39	All Levels	4	2016-05-02 19:26:48	Business Finance
506568	Create A Business From Home Trading Stocks Today In 2017	TRUE	75	10149	83	16	All Levels	2	2015-05-26 17:25:46	Business Finance
1020760	Introduction to Accounting : Mastering Financial Statements	TRUE	50	1916	38	23	Beginner Level	1	2016-12-05 22:14:17	Business Finance

Fig 1: Dataset after cleaning

V. RESULTS

1) Number of courses per category

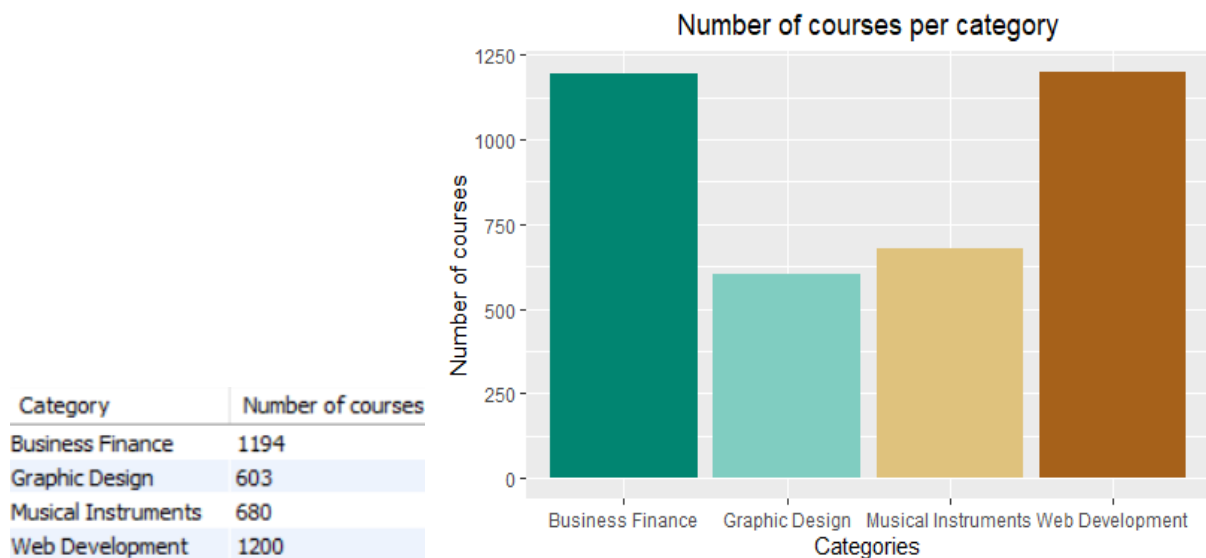| Category	Number of courses
Business Finance	1194
Graphic Design	603
Musical Instruments	680
Web Development	1200

Fig 2: Number of courses per category

The above bar graph will show which category has the greatest number of courses. It was obvious that Business & Finance and Web Development had the most courses. This might be due to a greater requirement for financial knowledge and awareness. There are various courses available on Udemy about tax preparation, asset diversification, and money management. Due to the diversity of learning to code by following along on one's own setup for web development, it may be easier for people nowadays to learn to code and create without needing to enroll Computer Science degree. Part-time work for programmers and developers is also widely available on the Internet and has emerged as one of the most promising occupations. I also find it amazing that there were more musical courses offered back then than graphic design courses.

2) Number of subscribers per Category

We can't understand the number of subscribers based on the number of courses per category since a category may have 200 courses with 10 viewers each and the course count would be more than another category with only 10 courses but 200 viewers each. To deal with this, I created a bar graph visualization that shows the number of subscribers by category.
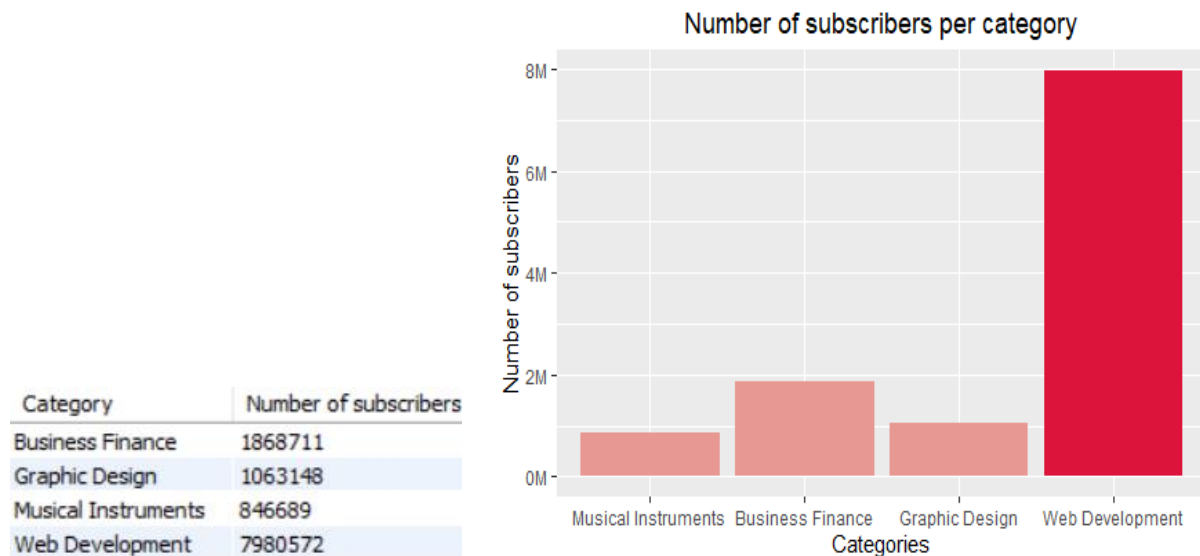

Category	Number of subscribers
Business Finance	1868711
Graphic Design	1063148
Musical Instruments	846689
Web Development	7980572

Fig 3: Number of subscribers per category

The accompanying bar graph provides further information about which category was the most popular on Udemy back then. Web Development has much more subscribers than the other categories. I believe this is due to the fact that the majority of students in today's culture are interested in pursuing professions in software development disciplines, and there is a high demand in the market for Web developers. As a result, Udemy is offering more Web development courses, and users are picking them.

3) Average price of the courses for all the categories

The cost of the course is a significant consideration for students who can only afford a limited budget. So, in general, online learning platforms keep the pricing dependent on the students in order to provide them with the finest information for the least amount of money. Now we'll look at the average pricing of courses across all categories.

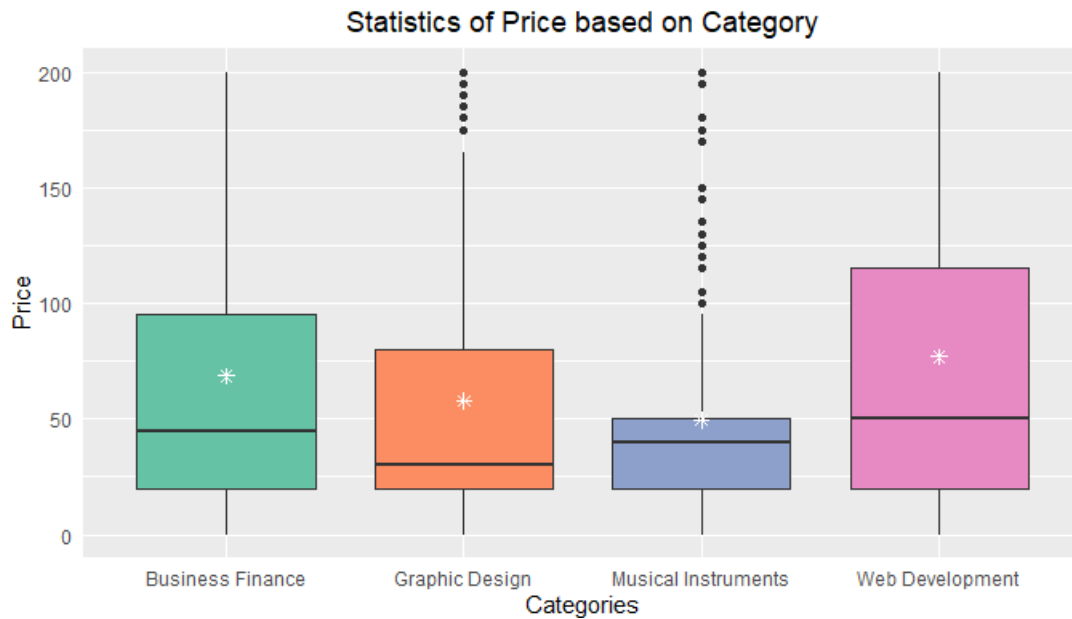| Business Finance		Graphic Design		Musical Instruments		Web Development	
count	1194.00	count	603.00	count	680.00	count	1200.00
mean	68.59	mean	57.83	mean	49.56	mean	77.03
std	64.39	std	57.21	std	41.19	std	66.08
min	0.00	min	0.00	min	0.00	min	0.00
25%	20.00	25%	20.00	25%	20.00	25%	20.00
50%	45.00	50%	30.00	50%	40.00	50%	50.00
75%	95.00	75%	80.00	75%	50.00	75%	115.00
max	200.00	max	200.00	max	200.00	max	200.00

Fig 4: Statistics of price based on category

Here's a boxplot illustrating the pricing spread per course type. So, based on the plot, we can conclude that the maximum price value for all categories is the same, which is $200 USD. However, the graph shows that many courses in the Web Development or Business & Finance categories will have that price, but only a few courses in the Graphic Design and Musical Instruments categories will have that price.

Even though Business & Finance and Web development have the same number of courses, the average price of Web development courses ($77.03) is higher than that of Business & Finance courses ($68.59). Because there isn't much demand for Musical Instruments courses, the average price ($49.56) is lower when compared to other categories. Graphic design is in the third position, with an average price of $57.83.
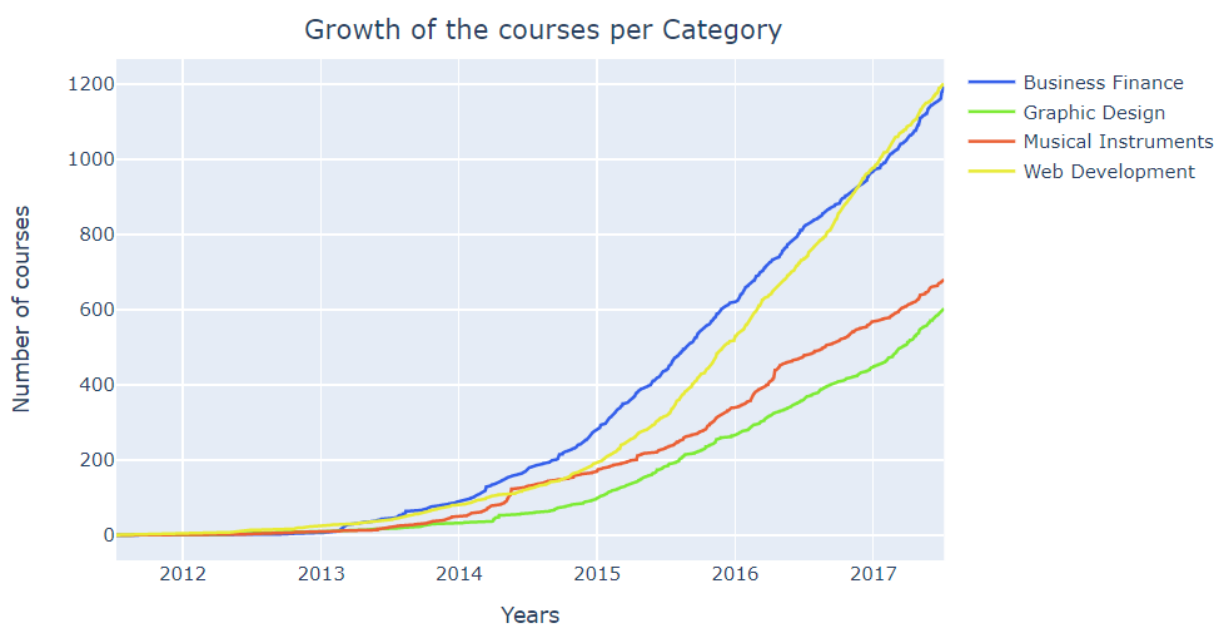
4) Growth of the courses per Category

Fig 5: Number of courses per category

The graph above depicts the rise of courses from 2012 to 2017. Udemy began to develop rapidly in 2013. When compared to other categories, Business Finance and Web development have grown substantially since 2015. Because there are several sorts of business and finance courses to study and teach, such as marketing, entrepreneurship, digital marketing, and so on, as well as numerous technologies, frameworks, and computer languages to learn and teach in web development. Musical instrument and graphic design classes, on the other hand, have a restricted number of courses and teachers. Another reason is that the market's need for Business Finance and Software development is extremely strong and never ends.

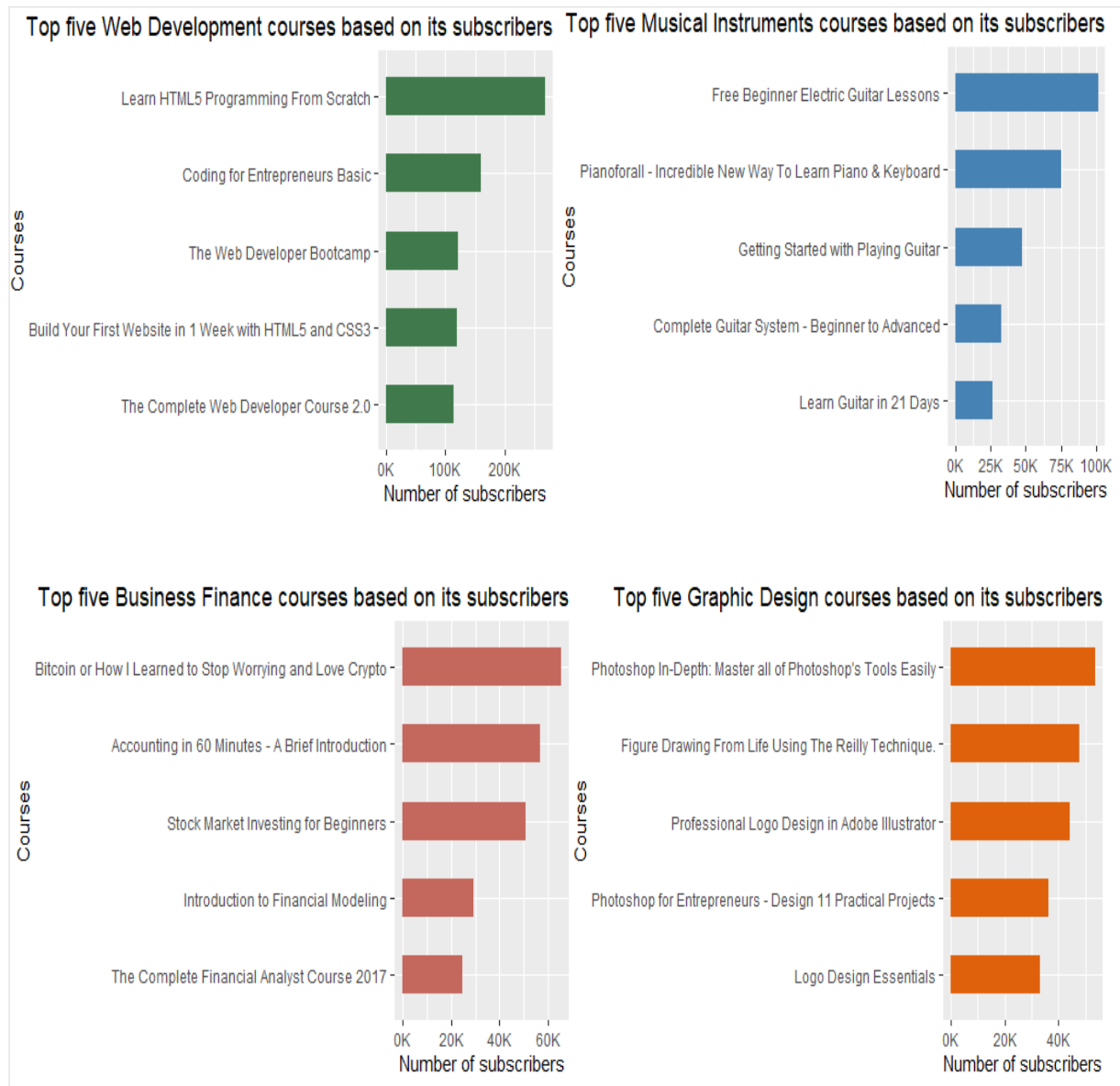## 5) Courses that have the most subscribers in each category

Fig 6: Top five courses that have the most subscribers in each category

The graph above depicts the top five courses in each category with the most subscribers. The course 'Learn HTML5 Programming From Scratch', which is connected to HTML, has the most subscribers (268K) in Web development, whereas the course 'Free Beginner Electric Guitar Lessons' has the most subscribers in Musical Instruments (101K). The course 'Bitcoin or How I Learned to Stop Worrying and Love Crypto' is the first in Business Finance, with 65K

members. Finally, the Graphic Design course 'Photoshop in Depth: Master All of Photoshop's Tools Easily' has more subscribers (53K).

6) Courses that have the most reviews in each category

In general, courses that have more subscribers should have more reviews. To check this I did a bar visualization for the courses that have the most reviews for all the categories. The following figure shows the top five courses that have the most reviews for all the categories.

Fig 7: Top five courses that have the most reviews in each category

The course 'The Web Developer Bootcamp' has the most reviews (27K) and is among the top five subscribers. The course 'Pianoforall - Incredible new approach to study Piano and Keyboard' has 7676 reviews in Musical Instruments and is ranked second on the subscription list. With 4397 reviews, the course 'Accounting in 60 Minutes - A brief Introduction' is top in Business Finance. With 2379 reviews, the course 'The Ultimate Drawing Course - Beginner to Advanced' is top in Graphic Design.

If you notice, all the courses with the most subscribers are not at the top of the list of courses with the most reviews. As a result, we may deduce that the reviews are not entirely dependent on subscribers.

7) Most expensive courses among all the categories

The table below shows the most expensive courses among all the categories of Udemy. As the maximum price of a course on Udemy is 200 USD (By referring to 3rd research question), there are a lot of courses on Udemy with that price. So, I filtered the courses with publication dates to show the top five expensive courses.

course_id	course_title	price	published_timestamp
12975	Introduction to Java Programming for Online Le...	200	2013-04-17 02:55:10
22825	Beginner Guitar Method	200	2012-10-31 04:13:51
30261	Kids Coding -Introduction to HTML, CSS and Jav...	200	2012-11-26 20:00:40
50528	The Complete Wordpress Plugin Development Cour...	200	2013-05-29 15:09:48
56513	Social Network Web Development	200	2013-09-03 06:29:24

Table 1: Most expensive courses among all the categories

8) Number of paid/free courses

We've seen which courses are the most expensive in each category, but pricing is a crucial consideration when choosing a course. We'll check it out using the pie chart representation below.
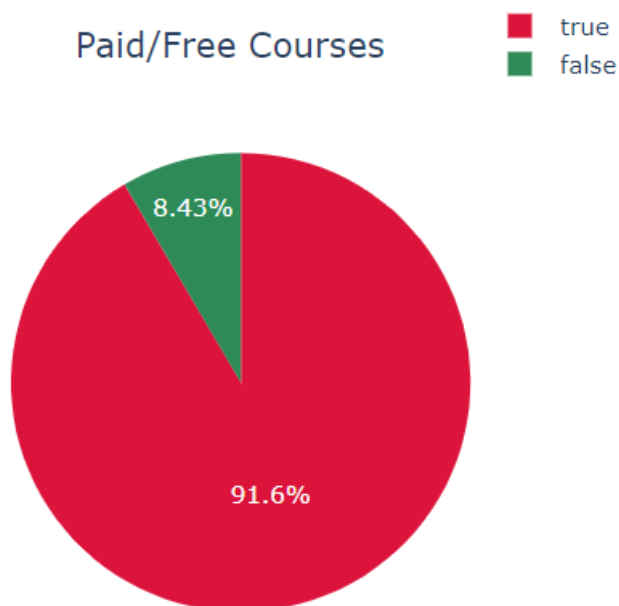

Fig 8: Paid and Free courses

Most of the courses on Udemy are paid. Even though just 8.4% of the courses are free, the number of members has increased dramatically year after year. This means users are giving importance to the material rather than the price.

I want to look at the correlation between the numerical variables to see which variables are more dependent on the others. I created a heatmap visualization to verify this. The graph below depicts the relationship between all of the numerical variables.
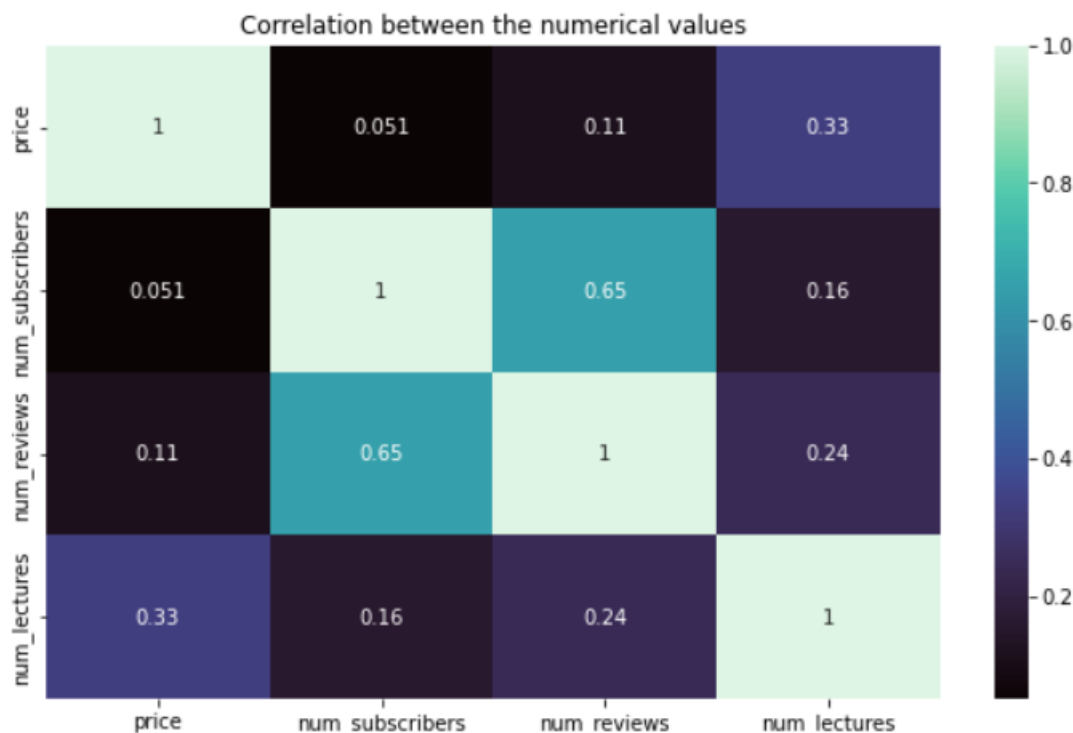

Fig 9: Correlation between the numerical values

Examine the above correlation heatmap for each numeric variable. The correlation between price and the number of subscribers (0.051) is quite modest. This might imply that most people are willing to pay a premium price as long as the material fits their requirements.

9) Most popular courses and when they are published

We know that pricing was unimportant and that the web development category had more subscribers than the others. More research is needed to identify the most defining elements. Before that, I wanted to discover which courses had the most subscribers. Let's take a look at the top five most popular Udemy courses in terms of membership.

course_id	course_title	num_subscribers	published_timestamp
41295	Learn HTML5 Programming From Scratch	268923	2013-02-14 07:03:41
59014	Coding for Entrepreneurs Basic	161029	2013-06-09 15:51:55
625204	The Web Developer Bootcamp	121584	2015-11-02 21:13:27
173548	Build Your First Website in 1 Week with HTML5 ...	120291	2014-04-08 16:21:30
764164	The Complete Web Developer Course 2.0	114512	2016-03-08 22:28:36

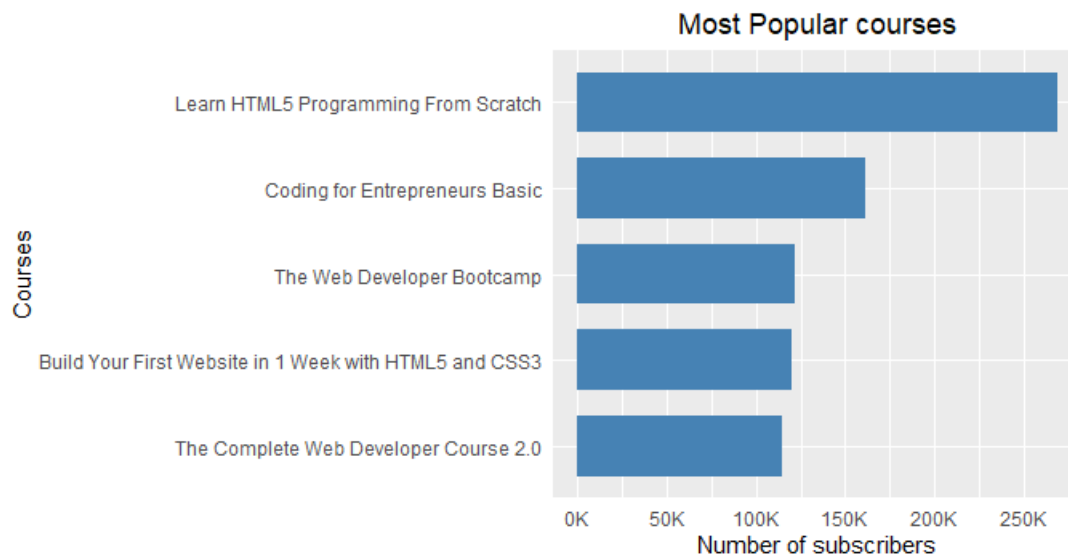Table 2: Most popular courses with the publication dates

Fig 10: Most popular courses

According to the graph above, the top five most popular courses are connected to the Web development area. The best course, "Learn HTML5 Programming From Scratch" was released in February 2013 with 268K subscribers. Because HTML is the foundation of Web development, the majority of users enrolled in this course.

10)Most Engaging courses and when they are published

Despite having fewer subscribers, there is a possibility of having more reviews for a few courses that consumers enjoy and submit ratings on. Reviews are used to determine which courses are the most engaging. We should be able to rank the most interesting ones based on the number of reviews.

course_id	course_title	num_reviews	published_timestamp
625204	The Web Developer Bootcamp	27445	2015-11-02 21:13:27
764164	The Complete Web Developer Course 2.0	22412	2016-03-08 22:28:36
756150	Angular 4 (formerly Angular 2) - The Complete ...	19649	2016-02-11 07:29:29
364426	JavaScript: Understanding the Weird Parts	16976	2015-03-12 23:19:54
705264	Modern React with Redux	15117	2015-12-28 19:20:21
289230	Learn and Understand AngularJS	11580	2014-09-24 00:57:25
461160	Learn and Understand NodeJS	11123	2015-08-31 21:07:48
41295	Learn HTML5 Programming From Scratch	8629	2013-02-14 07:03:41
719002	Angular 2 with TypeScript for Beginners: The P...	8341	2016-03-03 20:17:59
238934	Pianoforall - Incredible New Way To Learn Pian...	7676	2014-08-07 06:27:51

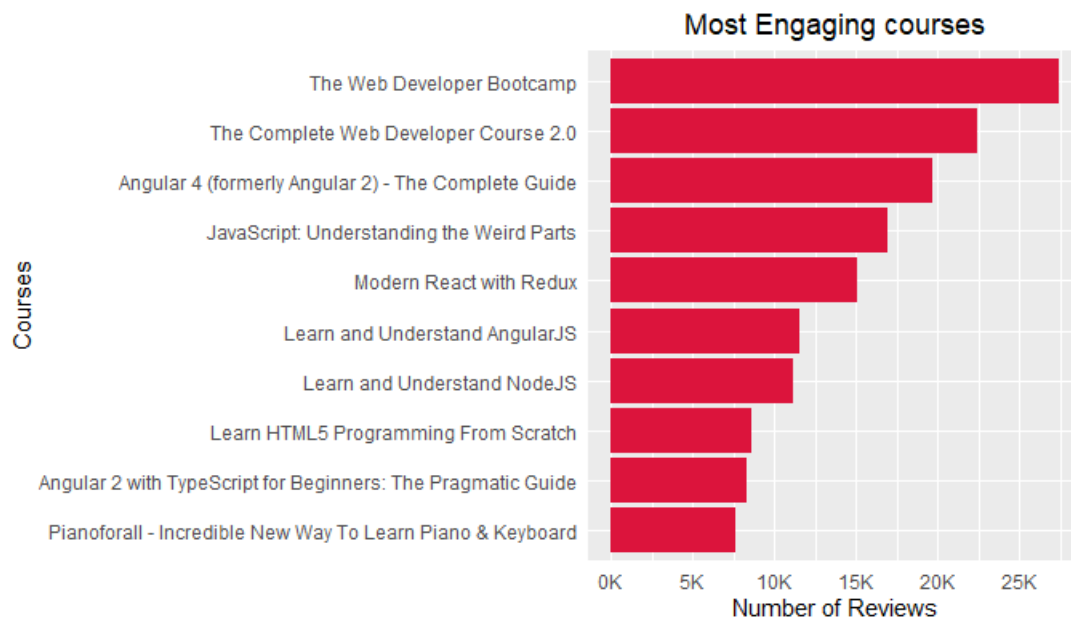Table 3: Most Engaging courses with the publication dates

Fig 11: Most Engaging courses

According to the above bar graph, the course "The Web Developer Bootcamp" has the most reviews (27445), although it ranks third on the list of most popular courses. In addition, the most popular course, "Learn HTML5 Programming From Scratch," with 268923 subscribers, is ranked eighth. Surprisingly, the Musical Instruments course "Pianoforall - Incredible New Way To Learn Piano" is among the top ten most interesting. With this, we might conclude that this course was previously quite popular.

11)Number of courses and their subscribers of all the levels

We will check the importance of the course levels by plotting the number of courses and their subscribers of all the courses.
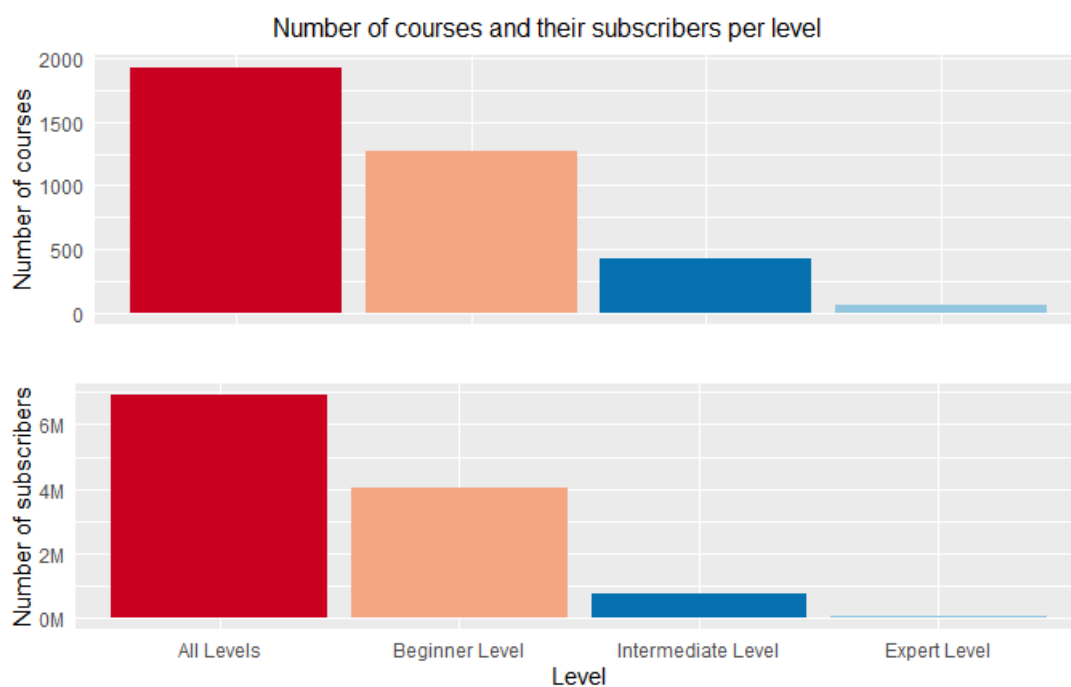

Fig 12: Number of courses and their subscribers per level

According to the graph above, most of the courses are available for 'All Levels,' followed by 'Beginner Levels,' 'Intermediate Levels,' and 'Expert Levels.' This means that the majority of users are given preference when selecting 'All Levels' courses. Furthermore, the 'Beginner Level' is significantly superior to the 'Intermediate Level' and 'Expert Level'. This explains why 'All Levels' and 'Beginner Level' courses are so densely packed on Udemy.

12) Average duration of the course based on the proficiency level

In general, 'Expert Level' courses last longer than the other levels, which is frequent. We will examine this using a box plot representation.

```
    All Levels           Beginner Level         Expert Level          Intermediate Level

count     1928.00      count      1270.00     count       58.00     count       421.00
mean         4.87      mean          3.09     mean         2.91     mean          3.73
std          7.31      std           4.05     std          2.53     std           4.40
min          0.13      min           0.45     min          0.52     min           0.50
25%          1.50      25%           1.00     25%          1.00     25%           1.50
50%          2.50      50%           2.00     50%          2.00     50%           2.00
75%          5.50      75%           3.50     75%          3.88     75%           4.50
max         76.50      max          78.50     max         12.50     max          31.50
```
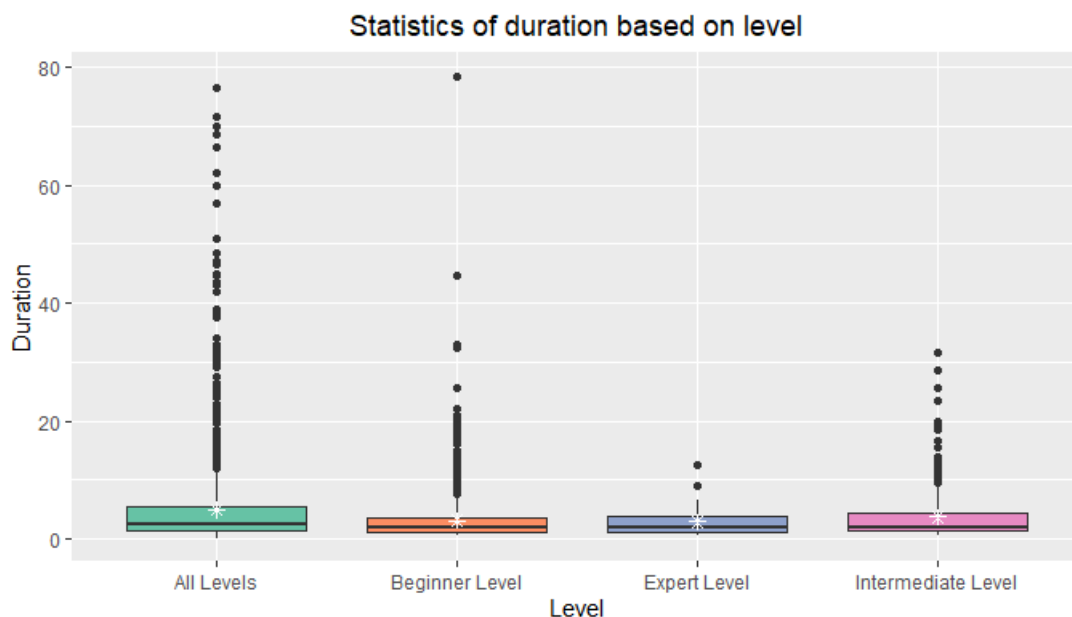

Fig 13: Statistics of course duration based on Level

According to the above box plot, the median course time appears to be consistent across all levels. And the average (mean) duration of the Expert Level is smaller (2.91 hours), followed by the Beginner Level (3.09 hrs), Intermediate Level (3.73 hrs), and All Levels (4.87 hrs). This indicates that my first theory appears to be incorrect. Expert-level courses appear to be generally shorter than lower-level ones. As a result, as your level rises, your predicted time to complete the course lowers.

13) Number of proficiency levels in each course category?

In general, the difficulty level of the courses varies depending on the category. The difficulty level may be low for certain categories and high for others. Professors also educate according to their standards in their respective categories. The total number of levels in each category is listed below.

Business &Finance

```
All Levels            695
Beginner Level        340
Intermediate Level    128
Expert Level           31
```

Graphic Design

```
All Levels            298
Beginner Level        243
Intermediate Level     57
Expert Level            5
```

Musical Instruments

```
Beginner Level        296
All Levels            276
Intermediate Level    101
Expert Level            7
```

Web development

```
All Levels            659
Beginner Level        391
Intermediate Level    135
Expert Level           15
```
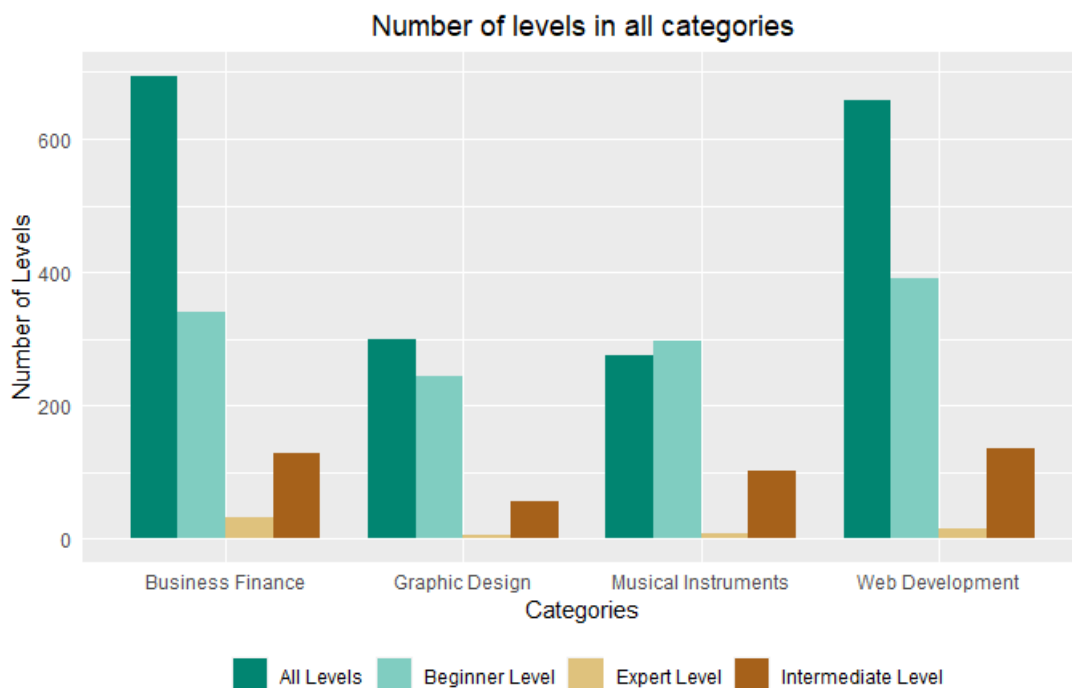

Fig 14: Number of levels in all categories

Based on the graph above, we can conclude that the number of 'All Levels' courses for Business & Finance and Web development is twice that of 'Beginner Level' courses. However, they are nearly identical to the Graphic Design and Musical Instruments courses. This demonstrates that the difficulty levels may differ depending on the category. As a result, for topics such as Business Finance and Web Development, teachers prioritize teaching All Levels of Concepts in a Single Course rather than breaking them down into Beginner, Intermediate, and Expert levels.

14) Relationship between variables

Now we'll look at the variables' relationships and how they affect one another. By doing so, we can finally identify what causes a course to get so many subscribers, so many reviews, and so on. To do this, I created a heatmap to determine the association.
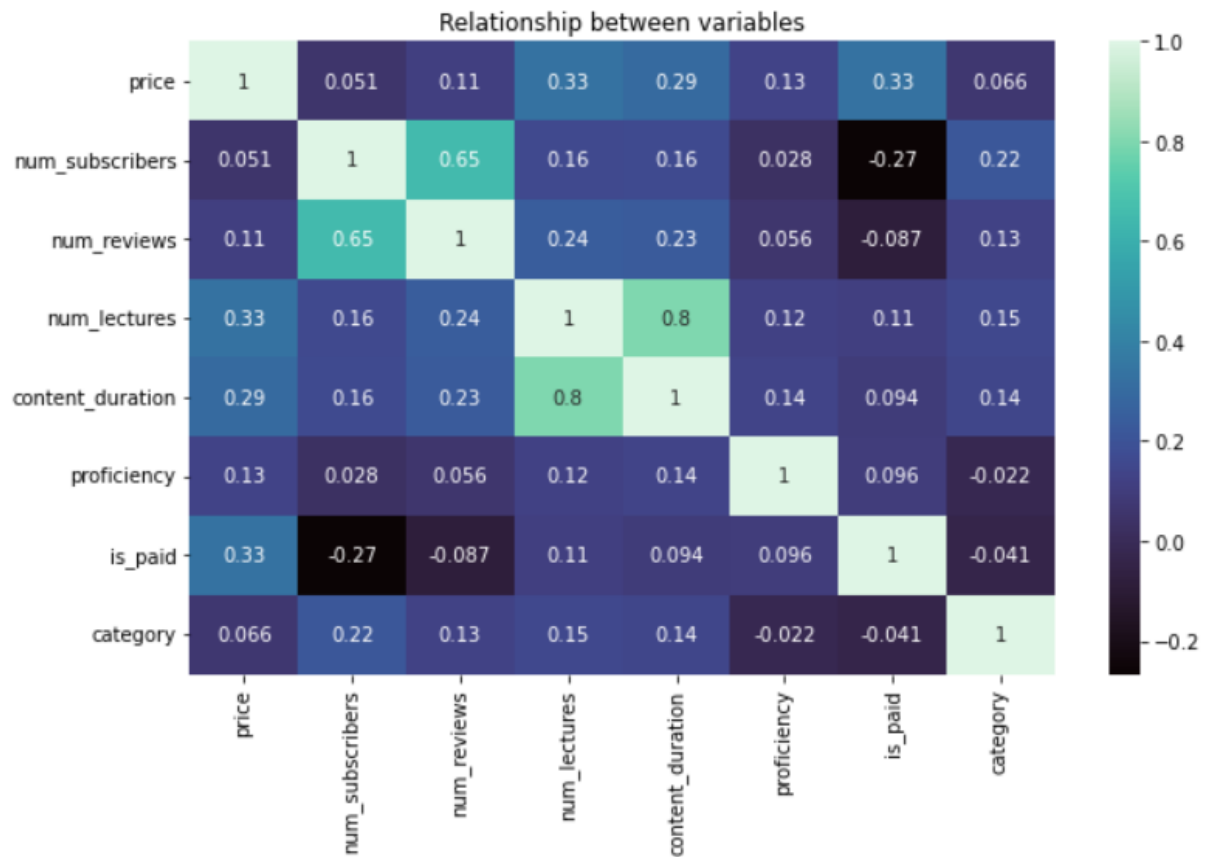

Fig 15: Relationship between the variables

This heatmap depicts the relationship of each variable. The relationship between content duration and the number of lectures is 0.8, which is clear since as the number of lectures grows, so does the duration. And, with a correlation of 0.65, the number of subscribers most closely corresponds with the number of reviews, indicating that the number of subscribers influences whether reviews rise or decrease. Furthermore, there is a -0.27 link between the number of subscribers and paid/free courses.

Random Forest Regression:

The most crucial aspect of every online course is the number of subscribers. We can observe that the number of subscribers has the strongest correlation with the number of reviews. However, owing to data restrictions, I was unable to determine which variable caused the rise in the number of subscribers. To find out, I decided to build a model with Random Forest regression.[9]

To measure the numerical columns for training data, I utilized Standard Scaler. The estimator was 600, the sample size was 0.5, and the random state was 77. The following figure shows the actual values and predicted values of the number of subscribers.
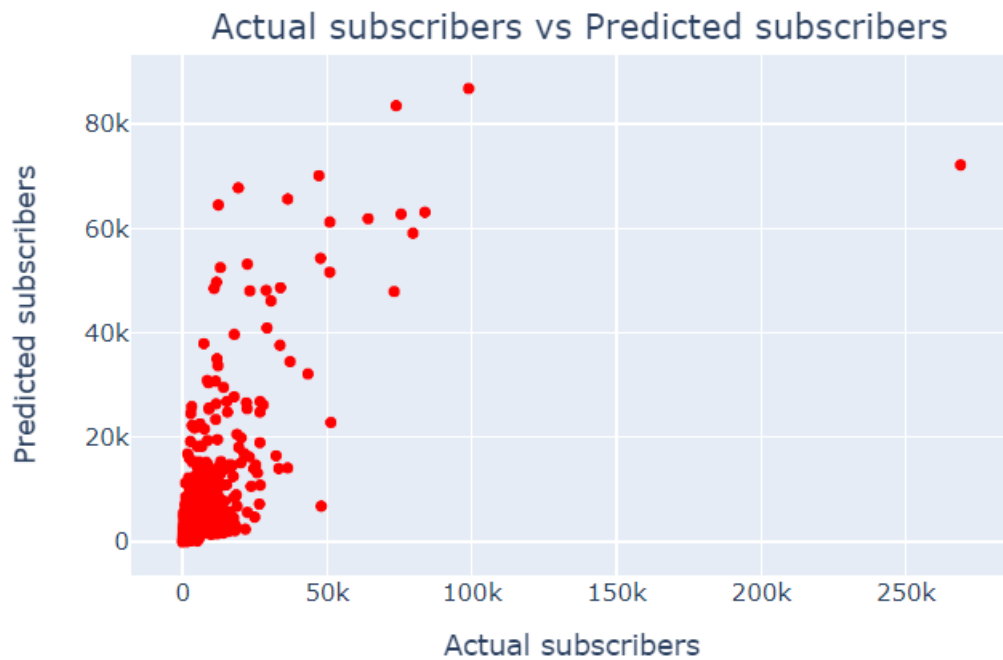
Fig 16: Actual subscribers vs Predicted subscribers

Looking at the figure, you can see that the model projected considerably higher values than the actual values, particularly in the 10K to 50K area on the X-axis. The MAE is 2127.94, whereas the MSE is 6565.21. I tried modifying the seed and test size several times, but the model's performance does not appear to be improving. We may also identify which features or variables are most important to the model.
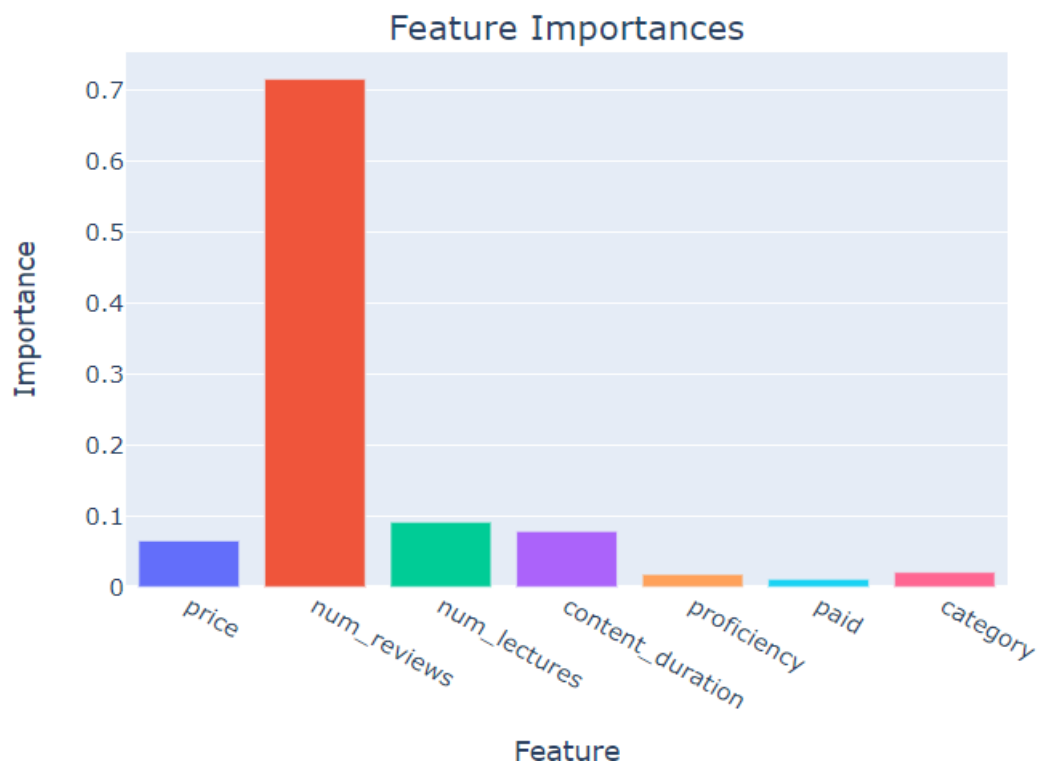

Fig 17: Importance of the variables for Random Forest Regression

The number of reviews per course appears to be the most relevant characteristic in Random Forest Regression for predicting the number of subscribers.

VI. SQL QUERIES

1) Display the first 10 rows of the dataset

```
SELECT * FROM UDEMY LIMIT 10;
```

course_id	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject
1070968	Ultimate Investment Banking Course	TRUE	200	2147	23	51	All Levels	1.5	2017-01-18 20:58:58	Business Finance
1113822	Complete GST Course & Certification - Grow Your CA Practice	TRUE	75	2792	923	274	All Levels	39	2017-03-09 16:34:20	Business Finance
1006314	Financial Modeling for Business Analysts and Consultants	TRUE	45	2174	74	51	Intermediate Level	2.5	2016-12-19 19:26:30	Business Finance
1210588	Beginner to Pro - Financial Analysis in Excel 2017	TRUE	95	2451	11	36	All Levels	3	2017-05-30 20:07:24	Business Finance
1011058	How To Maximize Your Profits Trading Options	TRUE	200	1276	45	26	Intermediate Level	2	2016-12-13 14:57:18	Business Finance
192870	Trading Penny Stocks: A Guide for All Levels In 2017	TRUE	150	9221	138	25	All Levels	3	2014-05-02 15:13:30	Business Finance
739964	Investing And Trading For Beginners: Mastering Price Charts	TRUE	65	1540	178	26	Beginner Level	1	2016-02-21 18:23:12	Business Finance
403100	Trading Stock Chart Patterns For Immediate, Explosive Gains	TRUE	95	2917	148	23	All Levels	2.5	2015-01-30 22:13:03	Business Finance
476268	Options Trading 3 : Advanced Stock Profit and Success Method	TRUE	195	5172	34	38	Expert Level	2.5	2015-05-28 00:14:03	Business Finance
1167710	The Only Investment Strategy You Need For Your Retirement	TRUE	200	827	14	15	All Levels	1	2017-04-18 18:13:32	Business Finance

2) Display the total number of courses per category

```
SELECT SUBJECT AS Category, COUNT(SUBJECT) AS `Number of courses` FROM UDEMY GROUP BY SUBJECT;
```

Category	Number of courses
Business Finance	1194
Graphic Design	603
Musical Instruments	680
Web Development	1200

3) Display the total number of subscribers per category

```
SELECT SUBJECT AS Category, SUM(NUM_SUBSCRIBERS) AS `Number of subscribers` FROM UDEMY
GROUP BY SUBJECT;
```

Category	Number of subscribers
Business Finance	1868711
Graphic Design	1063148
Musical Instruments	846689
Web Development	7980572

4) Display the average price of the courses per category

```
SELECT SUBJECT AS Category, ROUND(AVG(PRICE),1) AS `Average price` FROM UDEMY
GROUP BY SUBJECT ORDER BY PRICE DESC;
```

Category	Average price
Business Finance	68.6
Graphic Design	57.8
Web Development	77.0
Musical Instruments	49.6

5) Display the top five most popular courses

```
SELECT COURSE_ID AS `Course Id`, COURSE_TITLE AS `Course Title`,
NUM_SUBSCRIBERS AS `Subscribers`, PUBLISHED_TIMESTAMP AS `Date of published` FROM UDEMY
ORDER BY NUM_SUBSCRIBERS DESC LIMIT 5;
```

Course Id	Course Title	Subscribers	Date of published
41295	Learn HTML5 Programming From Scratch	268923	2013-02-14 07:03:41
59014	Coding for Entrepreneurs Basic	161029	2013-06-09 15:51:55
625204	The Web Developer Bootcamp	121584	2015-11-02 21:13:27
173548	Build Your First Website in 1 Week with HTML5 and CSS3	120291	2014-04-08 16:21:30
764164	The Complete Web Developer Course 2.0	114512	2016-03-08 22:28:36

VII. LIMITATIONS AND FURTHER RESEARCH

There were a few limitations in this study that were not addressed when working on this project. To begin with, this data is insufficient because it is limited to the years 2011 to 2017. We developed a model to estimate the number of subscribers using Random Forest regression, however the Mean square Error and Mean Absolute Error numbers that we obtained were excessively huge. As a result, we require additional data to correctly anticipate the model. As the maximum price of the course is limited to $200 USD, it is impossible to determine the expensive courses across all categories. Because there are numerous courses with the highest pricing, I sorted the top five courses in increasing order based on the published timestamp. Furthermore, there is no information on instructors in the data set, which is one of the key elements in enrolling in a course. And perhaps the instructors' ratings influence the number of subscribers to a course. As a result, further research is required to create the appropriate model to predict the number of subscribers by using the instructor's ratings and other features.

VIII. CONCLUSION

Overall, this project's data analysis was focused on the online learning platform Udemy, where users may choose from a range of courses to learn and enhance their skills based on their interests. This information is frequently presented in raw formats, and most people do not have the skill or time to investigate it. This study is quite useful for master's students who must gain and expand their knowledge and skills through self-learning using online learning platforms. Udemy has prioritized the Web development and Business Finance categories since students desire to pursue a profession in the software and business industries. The number of subscribers and reviews are mutually dependent (More subscribers more reviews and more reviews more subscribers). Furthermore, this study shows that if the course material satisfies their demands, users are unconcerned with the cost of the course.

References

[1]   Wyctor, "Text-multiclassification---PySpark," GitHub, 14 October 2022. [Online].
        Available: https://github.com/wyctorfogos/Text-multiclassification---PySpark.
        [Accessed 16 October 2022].

[2]   S. I. Bakhshi and A. Mohammad, "Legal Education through MOOCs: A Study of,"
        ResearchGate, July 2022. [Online]. Available:
        https://www.researchgate.net/publication/362430434_Legal_Education_through_MO
        OCs_A_Study_of_Select_International_Online_Platforms. [Accessed 30 October
        2022].

[3]   F. A. Khoirunnisa, V. S. Prabaningtyas, B. P. Prakoso and F. Y. Rukmana, "DATA
        COURSE ANALYSIS AND VISUALIZATION AT UDEMY," ResearchGate,
        December 2021. [Online]. Available:
        https://www.researchgate.net/publication/357093633_ANALISIS_DAN_VISUALISA
        SI_DATA_COURSE_DI_UDEMY. [Accessed 30 October 2022].

[4]   "Datacamp," Rdocumentation, [Online]. Available: https://www.rdocumentation.org/.
        [Accessed 29 November 2022].

[5]   "Python 3.11.0 documentation," Python, 04 December 2022. [Online]. Available:
        https://docs.python.org/3/. [Accessed 29 November 2022].

[6]   "ggplot2," RDocumentation, [Online]. Available:
        https://www.rdocumentation.org/packages/ggplot2/versions/3.4.0. [Accessed 29
        November 2022].

[7]   M. Waskom, "seaborn.heatmap," seaborn, 2022. [Online]. Available:
        https://seaborn.pydata.org/generated/seaborn.heatmap.html. [Accessed 29 November
        2022].

[8]   "MySQL Documentation," MySQL, 2022. [Online]. Available:
        https://dev.mysql.com/doc/. [Accessed 29 November 2022].

[9]   "sklearn.ensemble.RandomForestRegressor," scikit learn, 2022. [Online]. Available:
        https://scikit-
        learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.
        [Accessed 29 November 2022].

[10]   E. SZABO, "Online Platform Courses Between Education And Business," Ideas, 2021.
        [Online]. Available: https://ideas.repec.org/a/cbu/jrnlec/y2021v1p197-204.html.
        [Accessed 30 October 2022].