Data Analysis on Udemy dataset
Prepared by: Adekemi Kadri

Project Description
As a data analyst at Udemy, I have been given a task to present data on course revenue and identifying opportunities to increase revenues, as well as tracking course performance.

Data Design
To accomplish the task, I followed the steps below:

❖ **Dataset Description:** A total of four datasets in Excel format for four subjects (web development, graphic design, business finance, and musical instruments)were provided for this project, and the first step was to merge them together before analyzing. All four datasets contain similar columns with similar titles and contain details about the subjects. Each dataset contains the following details:
   I.   Course_id: This is a unique id number for each course
   II. Course_title: Title of course
   III. Url: The course's website
   IV. Price: Cost of course
   V.   Num_subscribers: Number of subscribers to the course
   VI. Num_reviews: NUmber of reviews for the course
   VII. Num_lectures: Number of lectures
   VIII. Level: Indicates the course level, ranging from beginner level, intermediate level, expert level, to all levels.
   IX. Rating: Values of the rating range from 0 to 1 on course
   X.   Content duration: Duration of course in hours
   XI. Published_timestamp: Timestamp of the date course was published on website
   XII. Subject: The subject category under which the course falls. There are four subject categoories: web development, corporate finance, graphic design, and musical instruments.

❖ **Data preparation and cleaning**: Before any data cleaning is done, the four datasets were merged once the data was collected and there was a total of 36,777 rows after merging.. Data cleaning is one of the most important processes in data analysis; in this project, it entails deleting duplicates, empty rows, improving the clarity and conciseness of the headers, and, ultimately, ensuring that the data is consistent. The subject title for web development was adjusted for consistency because it was discovered that it was different from other subject names. All of this data cleansing was done on Google Sheets.

❖ **Data Manipulation**: Another task to be done asides presenting the data on course revenue and identifying opportunities to increase revenues, as well as tracking course performance, it is required to get a list of top 20 most subscribed courses with the following information; their level, whether they are free or paid courses, whether they are any free beginner courses, duration of the course, and date they were published.

A certain amount of data manipulation is required in order to extract these details from the data. Before retrieving the list of the top 20 courses, I used Google Sheets' "left" function to extract the date that the courses were released from the timestamp since only the date and not the time is needed. Additionally, there isn't a column that indicates if a course is free or paid, so I added one by applying a logical function (the "if") to the pricing column to determine whether a course is free or paid.

After these data manipulations, the 20 most popular courses were extracted, and insights were produced from the entire dataset to address the project's primary objective using google sheets pivot table and tableau.

❖ **Insights from data analysis**

● Number of subscribers per subject: Web development is the subject with the most subscribers out of the four with almost 70% of the whole data(see figure 1 and 2), and it has tripled or more the subscribers of the other three. This is the first data-driven insight identified.
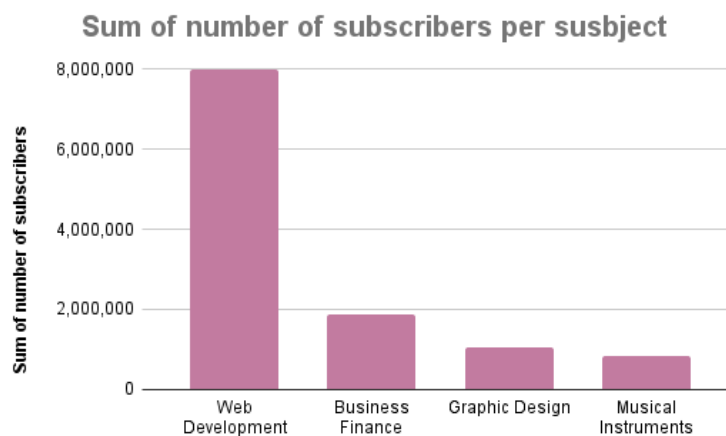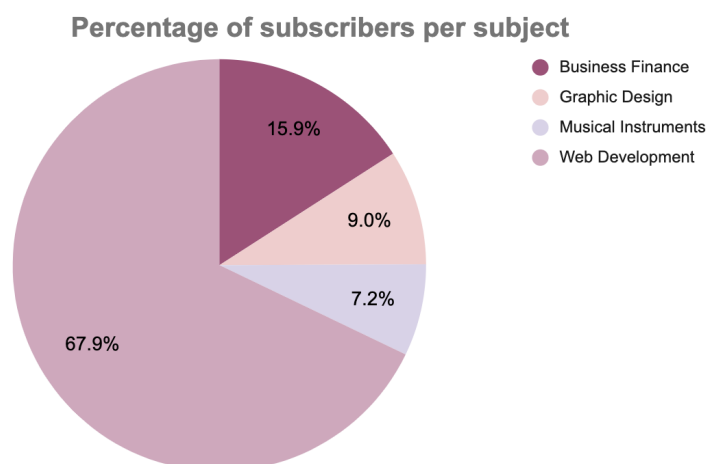

Figure 1: Number of subscribers per subject

Figure 2: Percentage of subscribers per subject

- Average cost per subject: Despite having the most expensive average cost, of its intermediate level courses, the web development field nonetheless has the most subscribers for its intermediate level courses as shown below(figure 3).ff
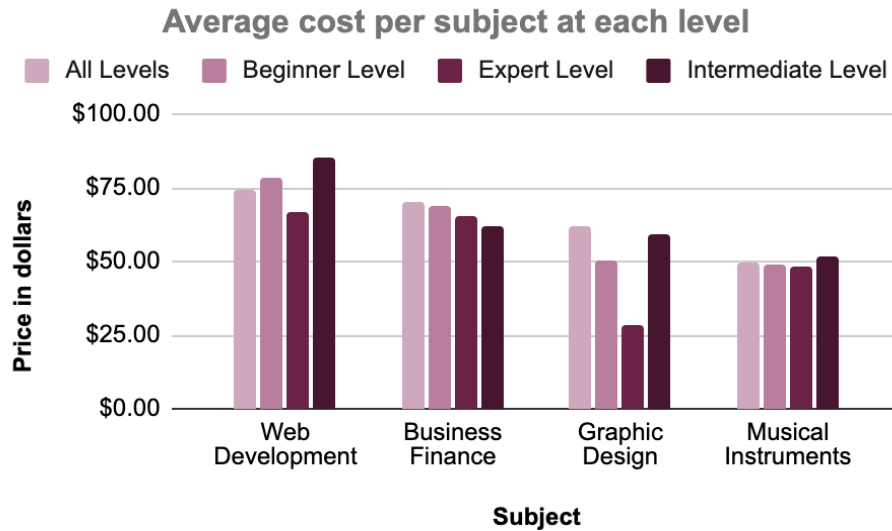
Average cost per subject at each level

Legend: All Levels | Beginner Level | Expert Level | Intermediate Level

Figure 3: Average cost of courses at different levels

- Ratio of free courses to paid courses: Web development continues to have the biggest number of subscribers on even paid courses when compared to others, with a significant disparity, web development has over 5 million subscribers of the paid courses and only approximately of 2.4 million of free courses, see figure 4.
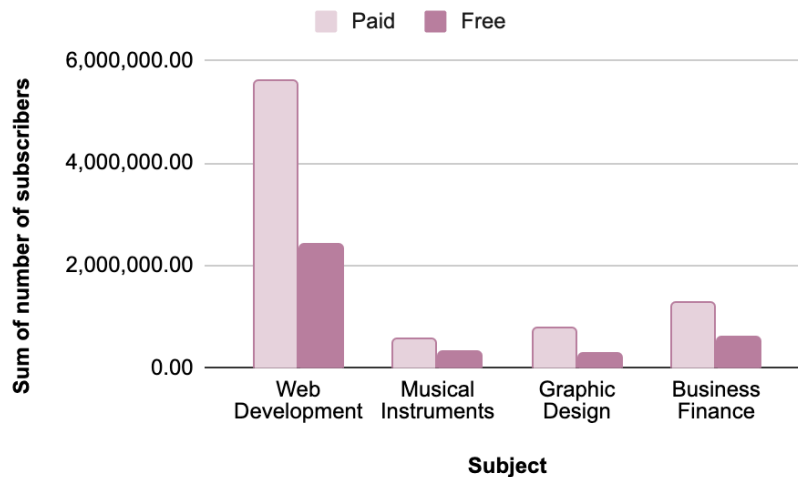
Legend: Paid | Free

Figure 4: Number of free and paid courses per subject

● Rating per subject: An intriguing finding from this data is that, in terms of average ratings, graphic design is rated highest, while web development, which is the main focus, is rated second-to-last in the pyramid, as seen below.
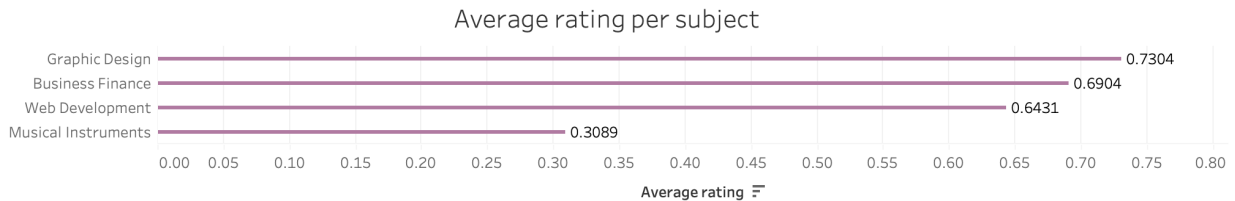
Average rating per subject

Graphic Design	0.7304
Business Finance	0.6904
Web Development	0.6431
Musical Instruments	0.3089

0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80

Average rating ⌶

Figure 5: Average Ratings per subject

● Number of subscribers yearly: Web development courses have experienced the highest number of subscribers compared to other subjects since 2011, growing year over year, but experiencing a fall between 2015 and 2017. It is important to recognise and comprehend the reasons for this recent decline in subscribers.
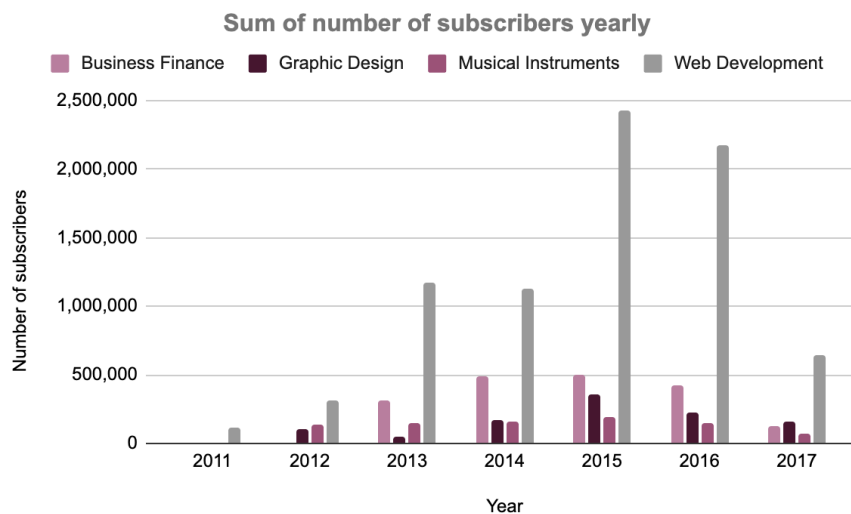
Sum of number of subscribers yearly

■ Business Finance   ■ Graphic Design   ■ Musical Instruments   ■ Web Development

Figure 6: Sum of subscribers yearly from each subject

● Sum of income yearly from the subjects: Since 2011, the total amount of income created annually has generally increased across all subjects, with Web development producing the most income, but in 2017, the income from each subject decreased, see table below and figure 7.

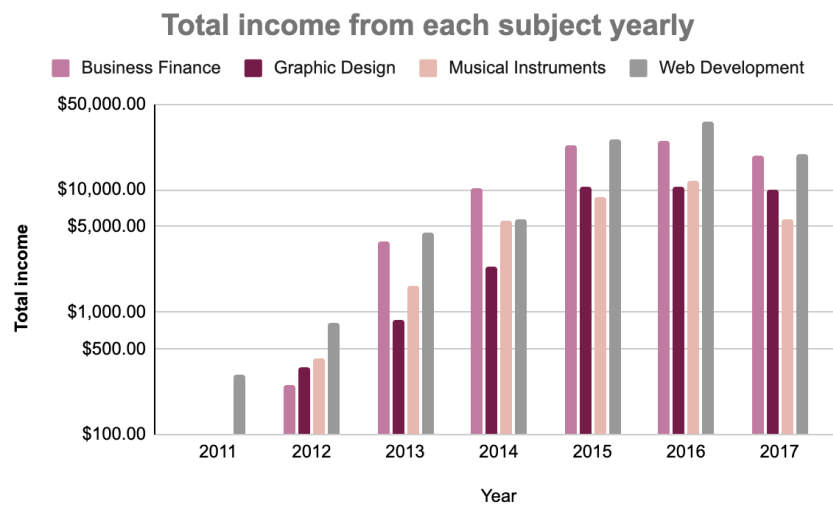| Year	Business Finance	Graphic Design	Musical Instruments	Web Development
2011				310
2012	255	350	415	815
2013	3825	875	1625	4460
2014	10205	2325	5545	5725
2015	23100	10505	8630	25595
2016	25405	10675	11735	36445
2017	19025	10120	5750	19325

Figure 6: Sum of income generated yearly from each subject

● Number of subscribers across all levels: Based on the analysis, the number of subscribers for all subjects decreased drastically after beginner level to intermediate level, see figure 7 below

Number of subscribers per level
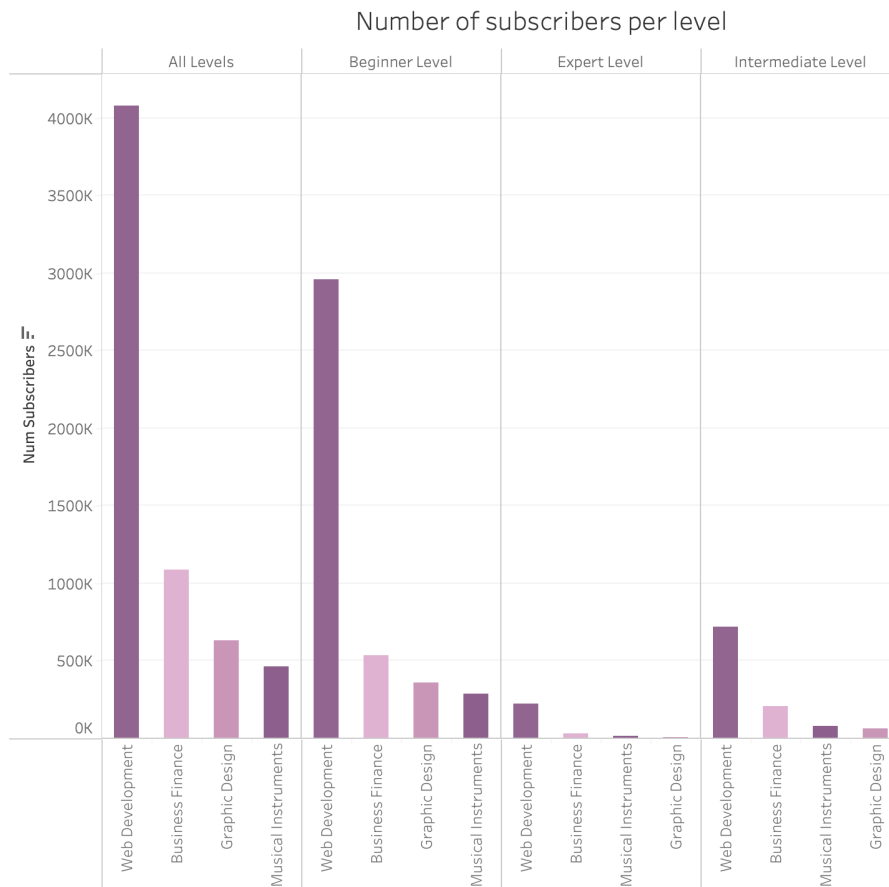
Conclusions and recommendations

There is no denying that web development has the potential to increase udemy's revenue because it is the most popular subject and has consistently produced the highest income over time, as seen from the analysis. However, it is important to note that web development did not receive high ratings from subscribers, which is something worth taking note of.

It is also important to consider other topics that could potentially increase revenue, such as corporate finance, which has consistently ranked below web development in most analyses and is second in terms of revenue created as well as having a higher rating.

It is important to be concerned when the number of subscribers decreases beyond the beginning level because it may be considered that the subscribers did not likely appreciate the beginner-level courses enough to wish to continue their subscriptions to the following level. This can be enhanced by making sure that beginner-level course content is thorough and pertinent enough to entice a member to continue paying for the next level

To see more insights generated, a tableau dashboard created can be found here.