**OVERVIEW OF THE PROJECT**

**INTRODUCTION**

This is a course project for python for data science. Under the scope of the course work, we are required to do an exploratory data analysis on the dataset provided to us using the techniques taught in the course. We will use Udemy Courses dataset for this project. This is a dataset which contains complete data for subjects (Business Finance, Graphic Design, Musical Instruments and Web Design) subscribed through 2011-2017.

**DESCRIPTION**

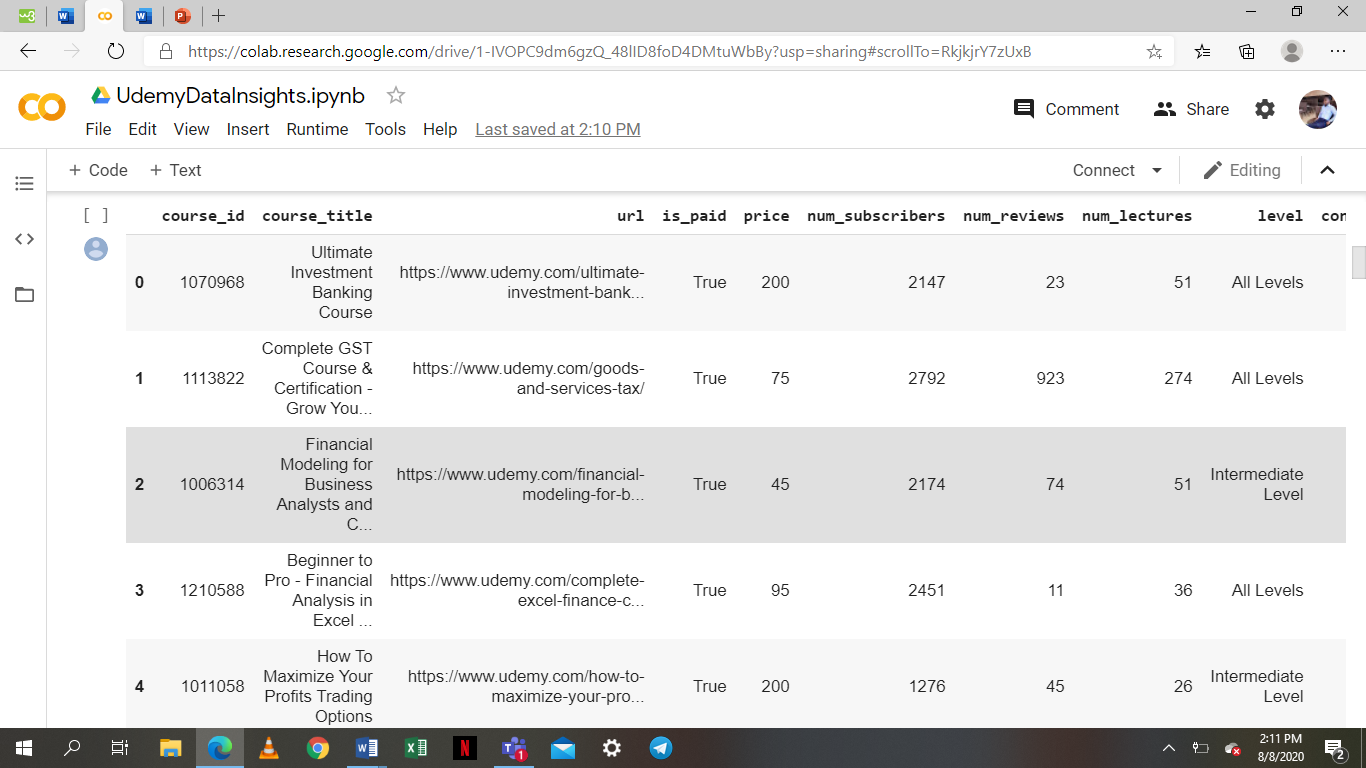
Udemy is a massive online open course (MOOC) platform that offers both free and paid courses. It is aimed at professional adults who want to add new skills to their resumes. Unlike academic MOOC programs that are driven by traditional collegiate coursework, Udemy provides a platform for experts of any kind to create courses which can be offered to the public, either at no charge or for a tuition fee. Udemy provides tools which enable users to create a course, promote it and earn money from student tuition charges.

In summary, our dataset contains a total of 3682 records with 12 features in a comma separated file. Each record in the file represent a course subscription. This dataset has different types of features such as categorical, continuous, numerical etc.

**DATASET**

This dataset has the following column names with explanation: Course ID as course\_id, Course Title as course\_title, Course URL as url, Whether the course is free or paid as is\_paid, Course Price as price, Number of subscribers as num\_subscribers, Number of reviews as num\_reviews, Number of lectures as num\_lectures, Course difficulty as level, Duration of all course materials as content\_duration, Date that the course was published as published\_timestamp and Course subject as subject.

Below is the screenshot of the sample data which includes only a few features and first 4 rows.



Below is the link to the Google Colab Notebook to see all the features in the dataset. <https://colab.research.google.com/drive/1-IVOPC9dm6gzQ_48lID8foD4DMtuWbBy?usp=sharing#scrollTo=RkjkjrY7zUxB>

**PROJECT SCOPE**

Our project scope is to run the exploratory data analysis using python and its libraries and to find business insights from our data. As per the recent studies, 10-20% of subscribers cancel their subscriptions every year. This is a huge risk for the investors who are funding Udemy. Investors require more comprehensive assessment of the course subscriptions to make a smart business decision.

**DATA CLEANING AND EXPLORATORY ANALYSIS**

The data used for this project is the structured data with few missing or null values. We will use the data cleaning or imputation techniques for the missing values imputation. Duplicate values are removed from the dataset. We also noticed some outliers in subject web development in relation to the number of subscribers. This was identified using box plots and scatter plots.

We removed some features which have missing values for about 20% of the data. In addition, we removed the Course URL column and the Course title column since the data doesn’t contain information that will be useful for data analysis.

After cleaning the data, we used python to run exploratory data analysis to find the business insights from the data. We extensively used python libraries; pandas, numpy, matplotlib and seaborn for the data analysis.

**BUSINESS PROBLEM**

As the years go by, the number of paid subscribers decreases. Possible reasons for this are:

1. There are several online platforms available.
2. There are more free courses available.
3. The courses available are too expensive.
4. The difficulty level of the courses is high i.e. for expert levels.
5. The number of course reviews keeps decreasing.
6. The duration of the course videos is long.

**HYPOTHESIS**

The number of the subscribers for paid Udemy courses are decreasing over the years.

This hypothesis was motivated by the assumption that there are several free courses available online now, and so subscribers are less interested in paying for courses they can get for free elsewhere. Hopefully, the data available from Udemy will give insights that will help them to improve their course structure.

**QUESTIONS**

To be able to test this hypothesis, a few questions were asked. This will help in gathering information for data analysis.

* Are the number of subscribers for free courses higher than those for higher priced courses?
* Are the number of subscribers for high level of difficulty courses higher than those courses of lower level of difficulty?
* Are the number of reviews for free courses higher than those for paid courses?
* Is the content duration for free courses higher than that of the paid courses?
* Are the number of lectures for paid courses more than that of the free courses?
* Are the number of subscribers for Business finance more than that of graphic design/musical instruments/web design?
* Were the number of subscribers per course increasing over the years?
* Was the price per course increasing over the years?

**VISUALIZATION AND INSIGHTS**

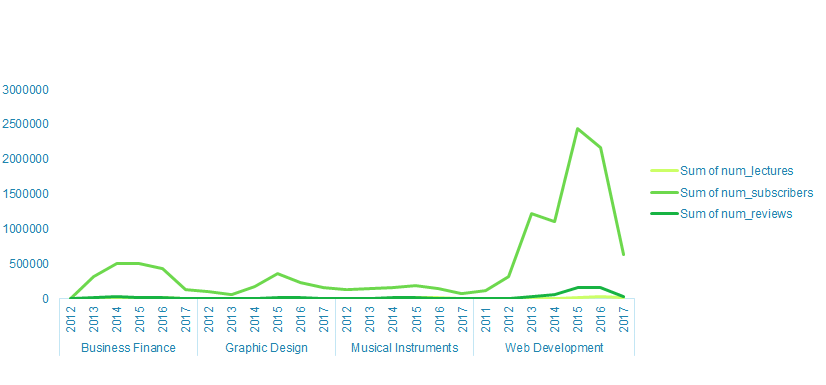
* Trends and Distributions

Trend in the number of subscribers for subjects from 2011-2017

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Business  Finance | Graphic  Design | Musical Instruments | Web Development | Total for the  year |
| 2011 |  |  |  | 119028 | 119028 |
| 2012 | 3620 | 100649 | 133635 | 317435 | 555339 |
| 2013 | 311664 | 50133 | 150224 | 1211417 | 1723438 |
| 2014 | 494626 | 174582 | 156152 | 1105049 | 1930406 |
| 2015 | 501858 | 352856 | 190368 | 2430242 | 3475342 |
| 2016 | 426647 | 229587 | 148748 | 2161662 | 2966644 |
| 2017 | 130299 | 155341 | 67562 | 635739 | 988941 |

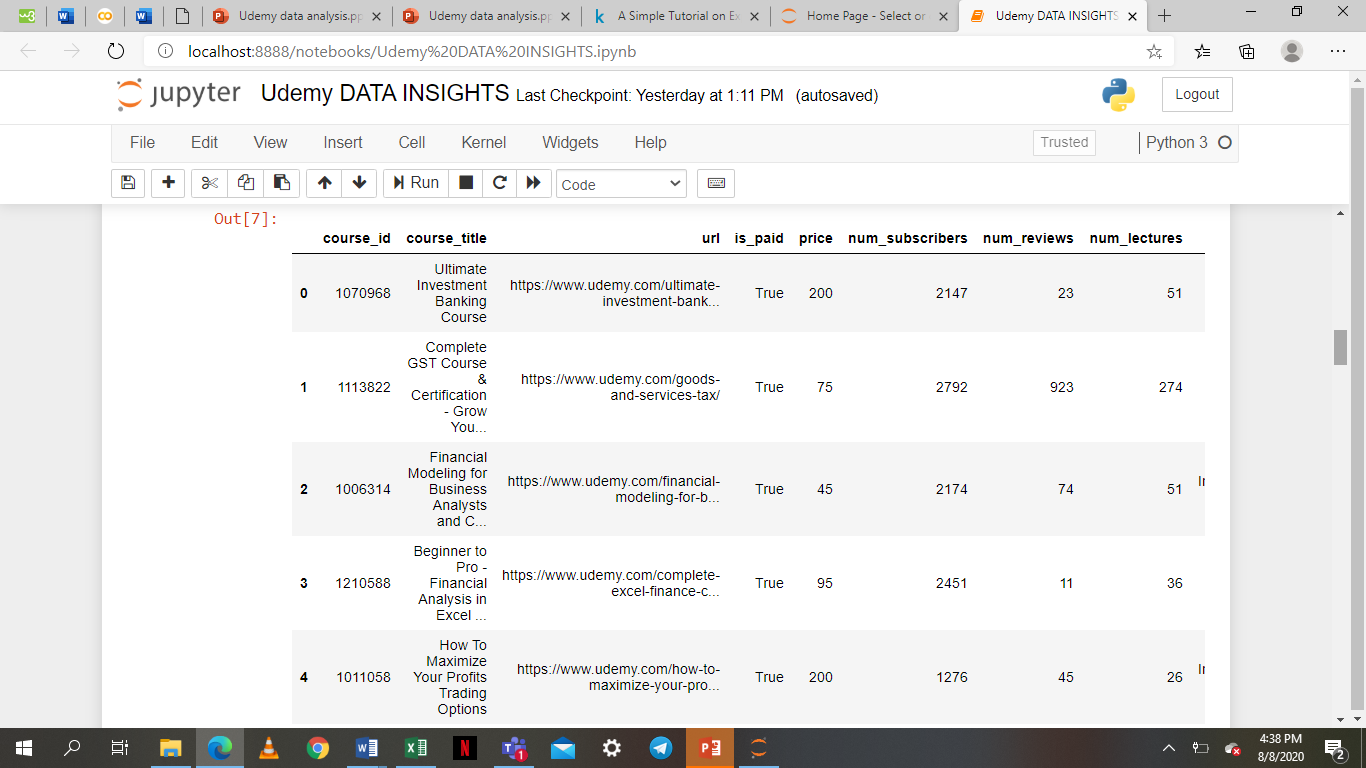
From the table, the number of paid subscribers decreases by 10-20% each year. More people subscribed to web development with a total number of subscriptions amounting to 7,980,572 through 2011-2017. Fewer people subscribed to musical instruments with total subscriptions amounting to 846,689 through 2011-2017.

**A graph showing the total number of lectures, subscribers and reviews per course**

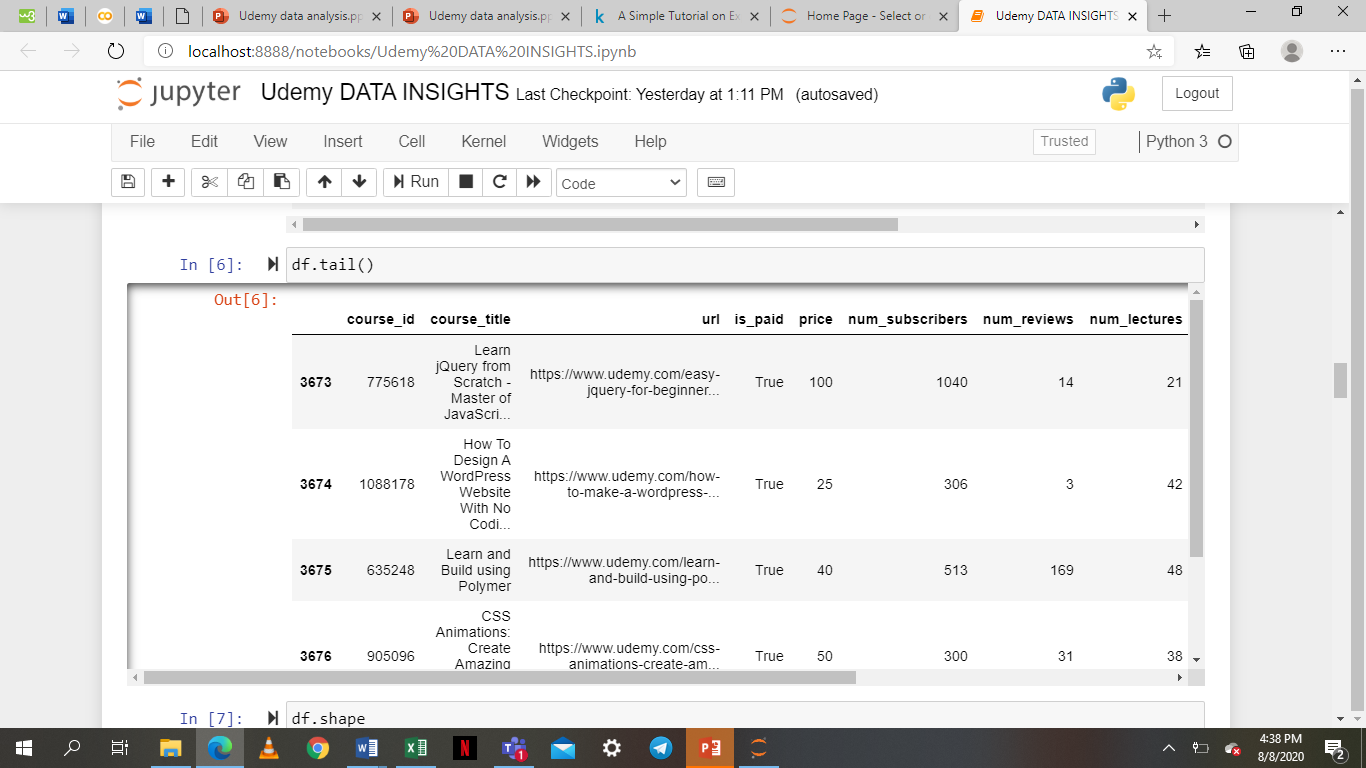


Now that we have got a general idea about our data set, it’s also a good idea to take a closer look at the data itself. With the help of the head() and tail() functions of the Pandas library, we can easily check out the first and last lines of the Data Frame (udemy\_data), respectively.

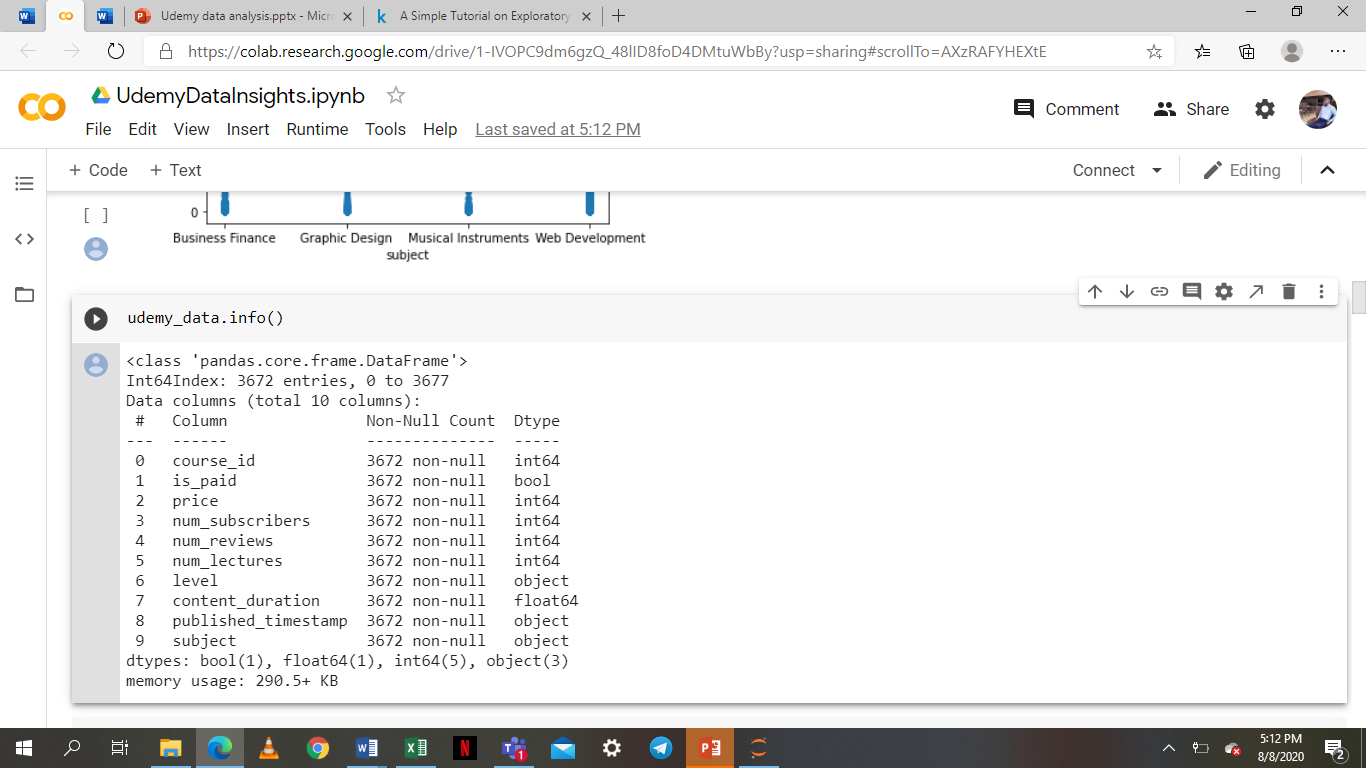
This shows the first 5 rows



This shows last 5 rows

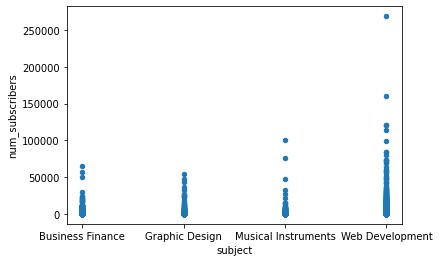


This gives a brief information about the data set. The information includes the data types and memory usage.



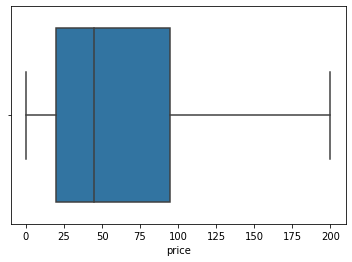
* Outliers

This is a scatter plot to determine if there are any outliers in the number of subscribers per subject.



From the scatter plot above, there were outliers in the subject web development in relation to the number of subscribers.

This is a plot to check if there are any outliers in price.



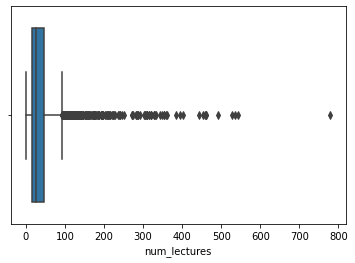
There were no outliers in the price.

This is a plot to check if there are any outliers in number of reviews.



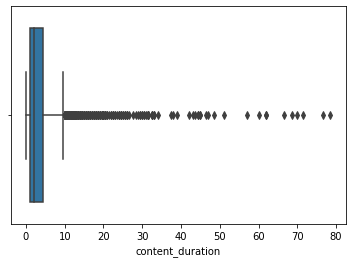
There were no outliers in the number of reviews.

This is a plot to check if there are any outliers in number of lectures.



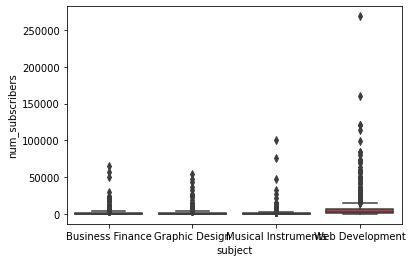
There were outliers in the number of lectures.

This is a plot to check if there are any outliers in content duration.



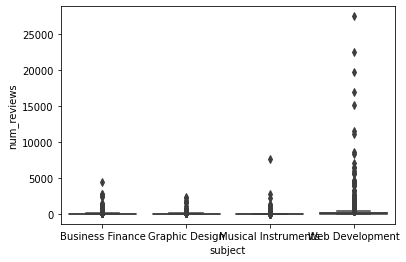
There were no outliers in the content duration.

This is a plot to check if there are any outliers in number of subscribers per subject.



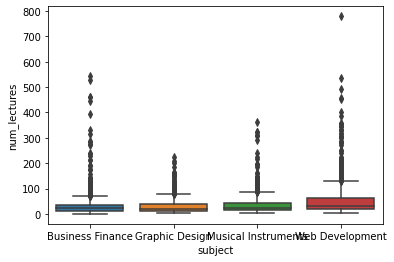
There were outliers in the number of subscribers with relation to the subject web development.

This is a plot to check if there any outliers in number of reviews per subject.



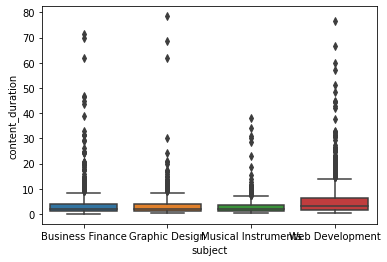
There were no outliers in the number of reviews.

This is a plot to check if there are any outliers in number of lectures per subject



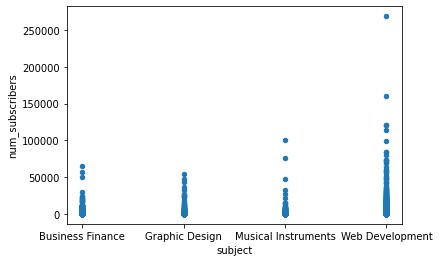
There were no outliers in the number of lectures.

This is a plot to check if there are any outliers in content duration per subject.



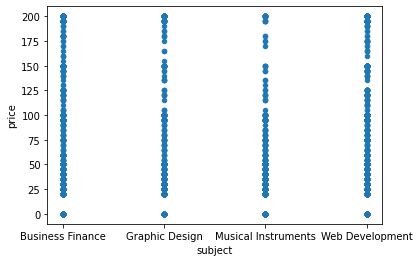
There were no outliers in the content duration.

This is a plot to check if there are any outliers in number of subscribers per subject.



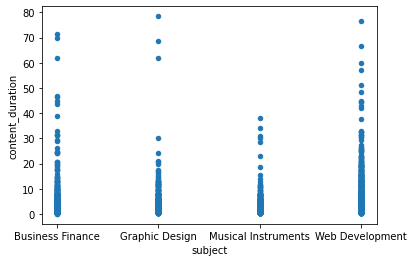
There were outliers in the subject web development in relation to the number of subscribers.

This is a plot to check if there are any outliers in subject per price.



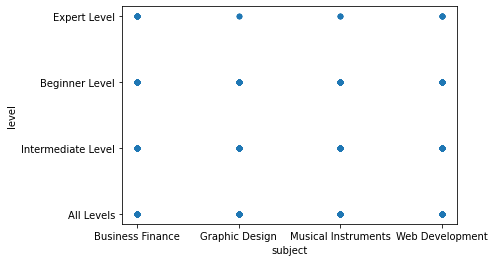
There were no outliers in relation to the subjects and the price.

This is a plot to check if there are any outliers in content duration per subject.



There were no outliers in relation to the content duration and the subject.

This is a plot to check if there are any outliers in level.



There were no outliers in relation to the level and the subject.

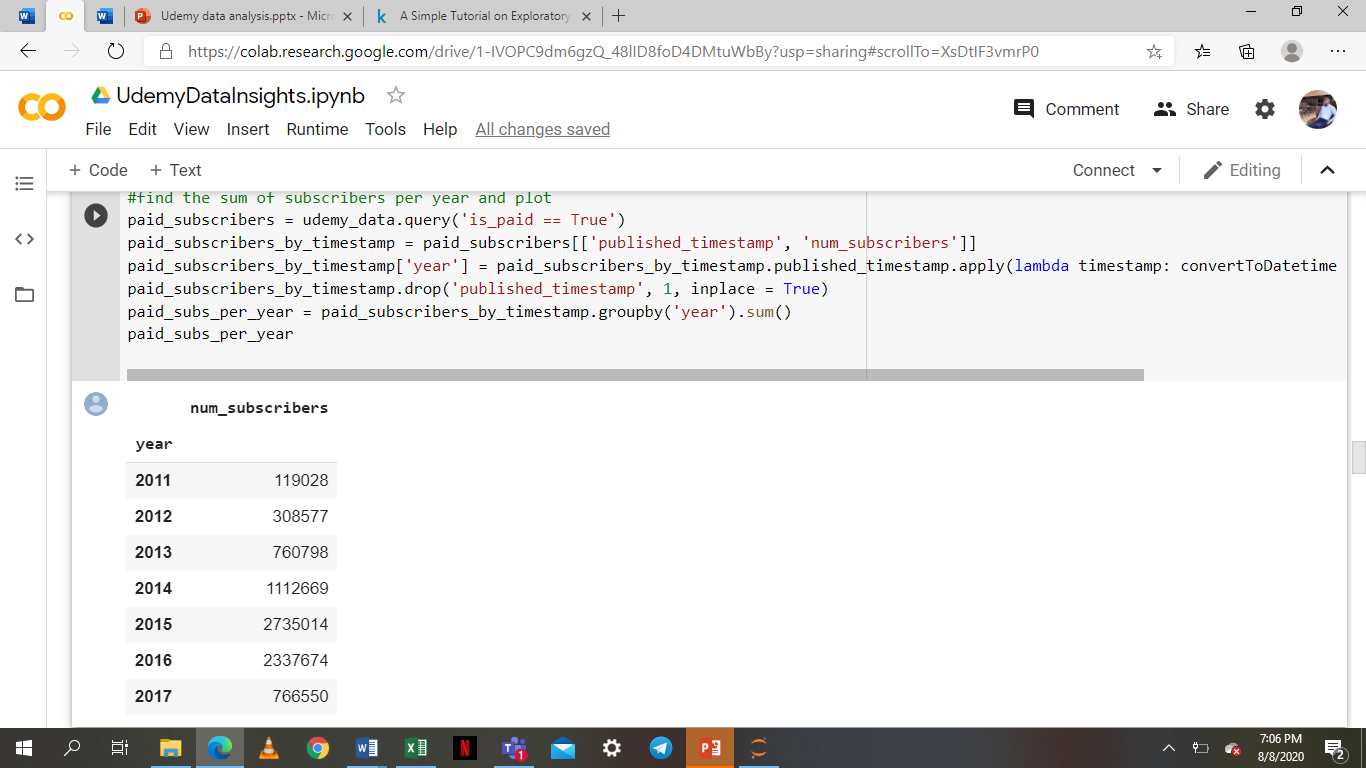
* Spread measurements

|  |  |
| --- | --- |
| PAID AND FREE COURSES | TOTAL |
| number of subscribers doing paid courses | 8140310 |
| number of subscribers doing free courses | 3575525 |
| number of reviews for paid courses | 442158 |
| number of reviews for free courses | 132038 |

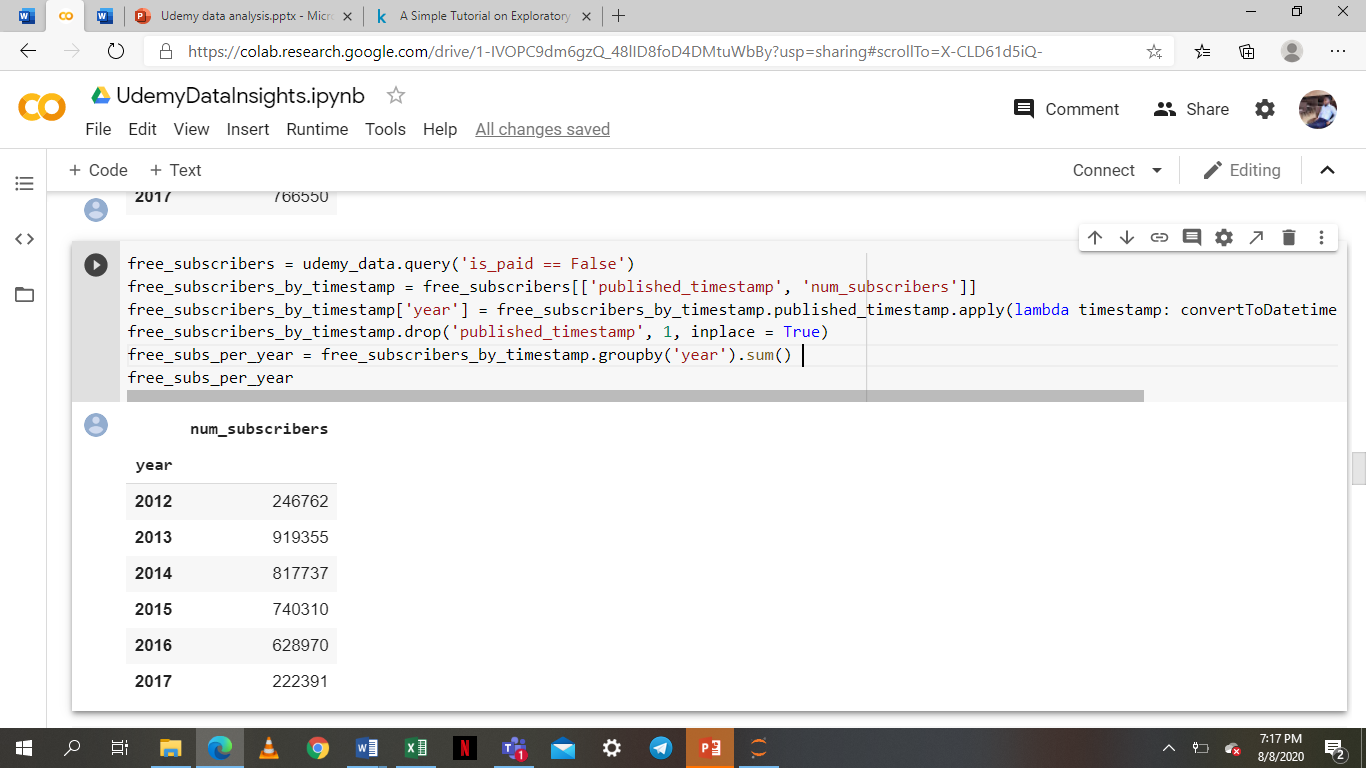
**DATA ANALYSIS**

* Hypothesis Testing

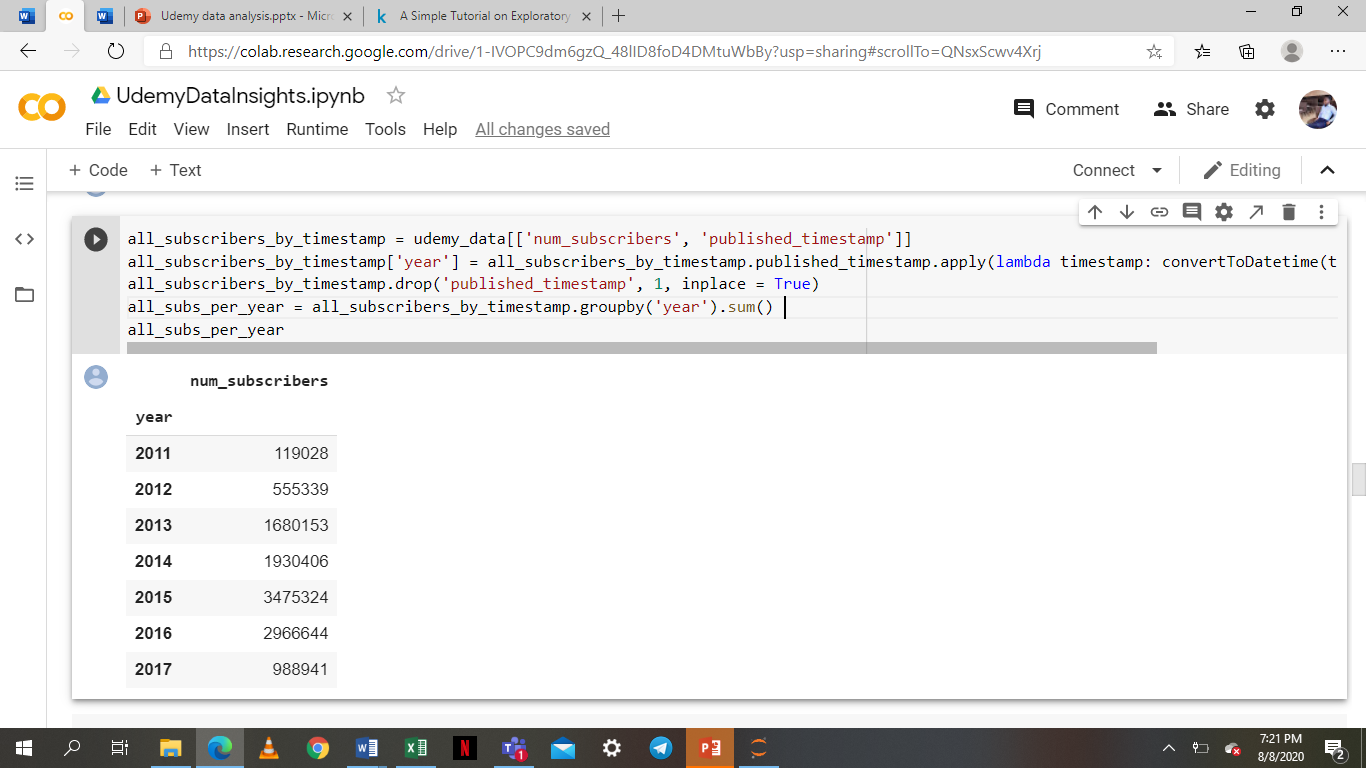
This is to find the sum of paid subscribers per year.



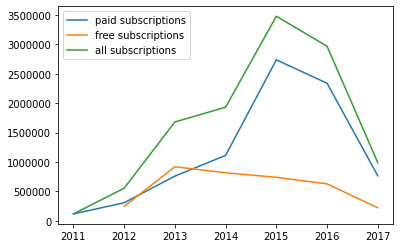
This is to find the sum of free subscribers per year.



This is to find the sum of all subscribers per year.



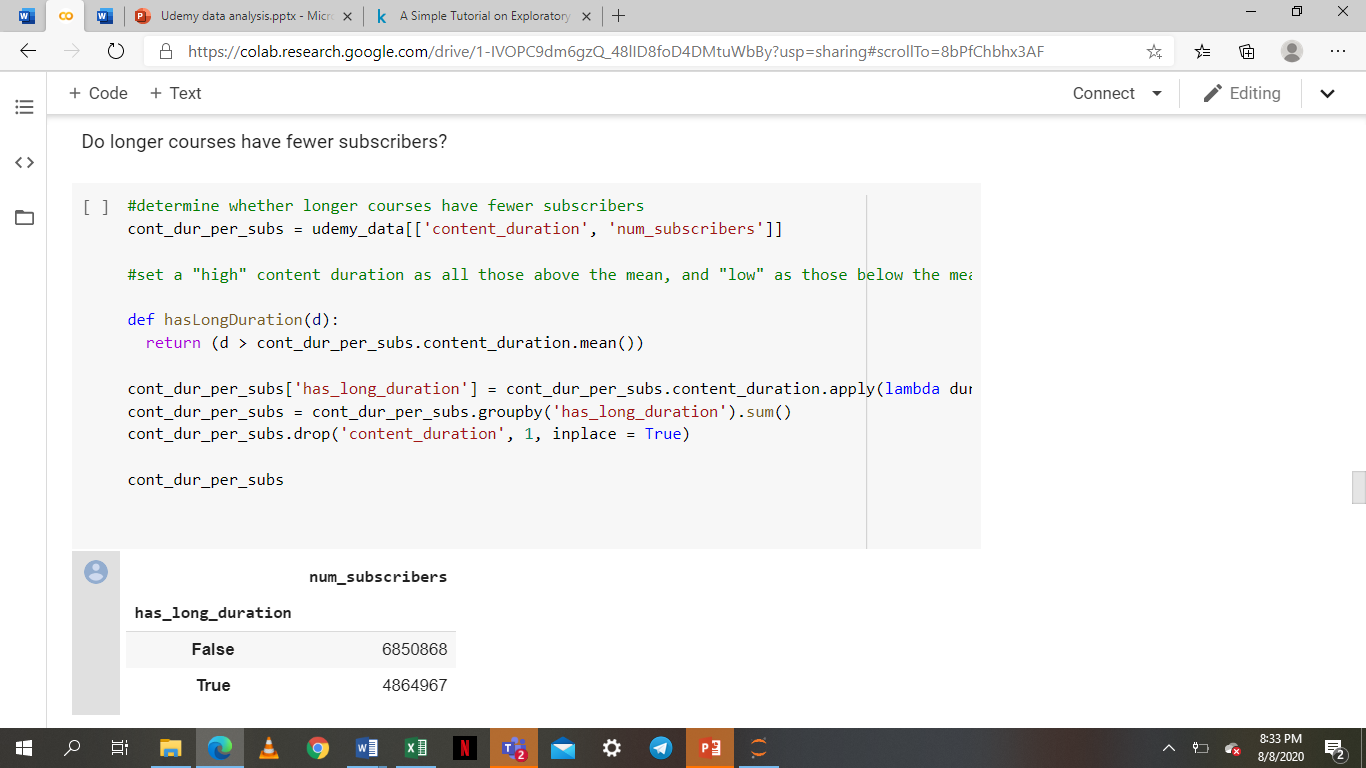
This is a plot of the number of subscribers per year.

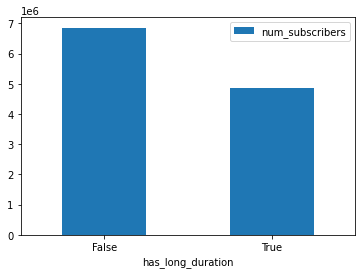


The plot indicates that the number of paid subscribers is decreasing over the years. This may be due to the fact that the overall number of subscribers are decreasing. However, it is noticeable that the number of subscribers taking free courses is far lower per year than the paid ones.

**Question 1**

Do longer courses have fewer subscribers?

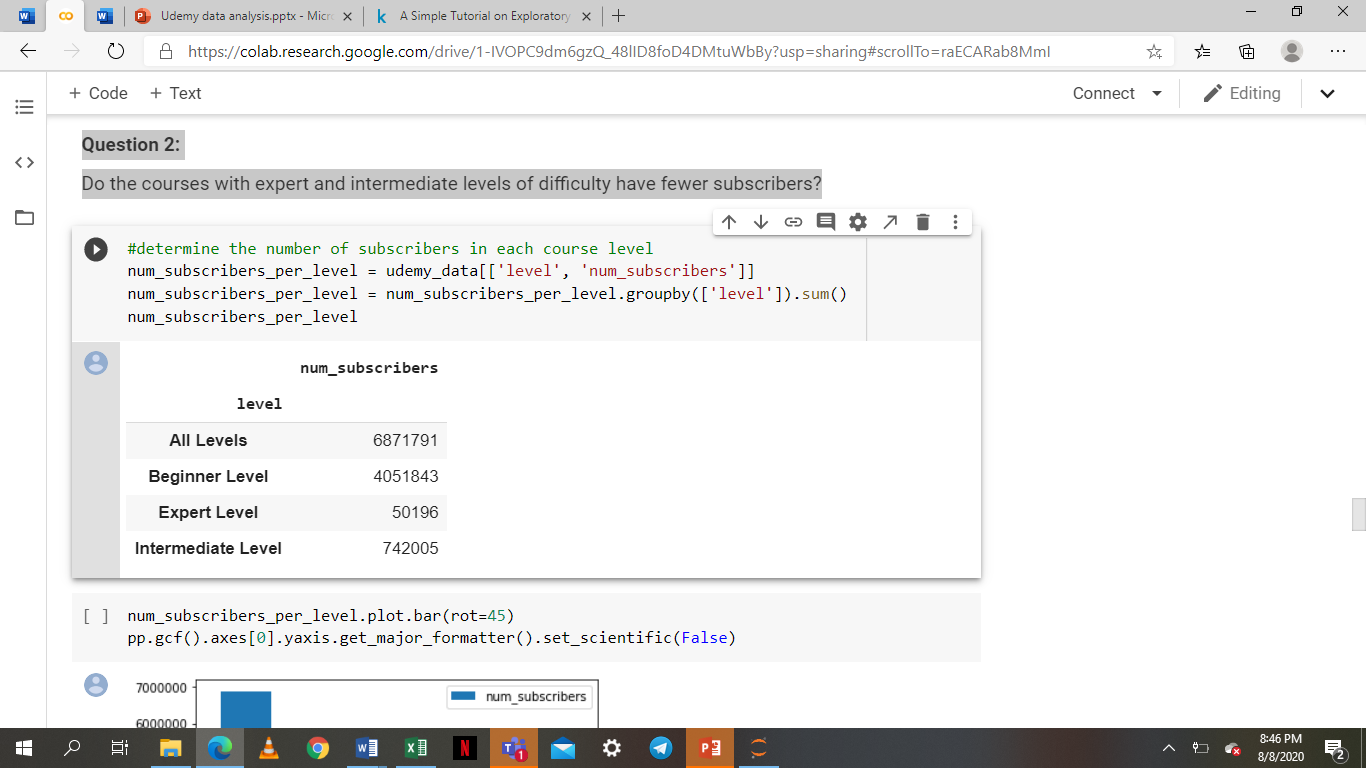


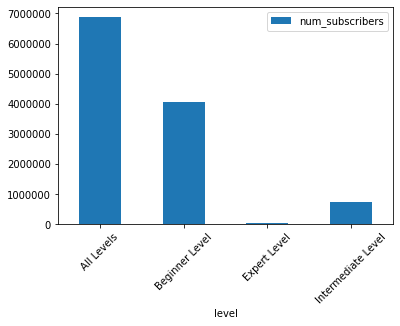


The plots show that courses with content duration above the mean have fewer subscribers. This may indicate that subscribers prefer shorter courses.

**Question 2**

Do the courses with expert and intermediate levels of difficulty have fewer subscribers?

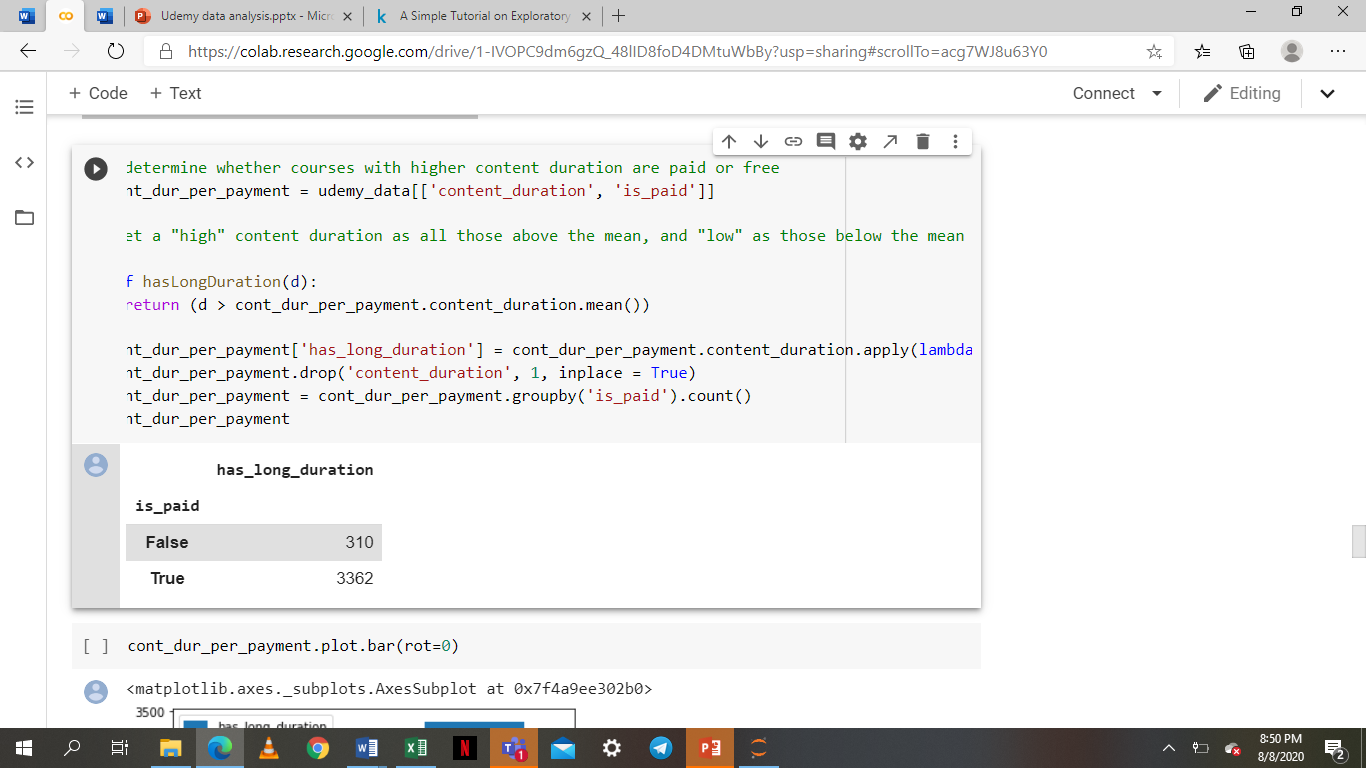


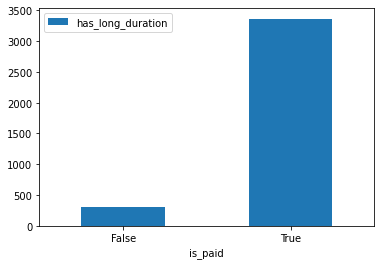


Expert level courses significantly have the least subscribers. Intermediate level courses also have less than half the number of subscribers that beginner level courses have. This indicates that more people subscribe to beginner level courses.

**Question 3:**

Are the paid courses longer than free courses?

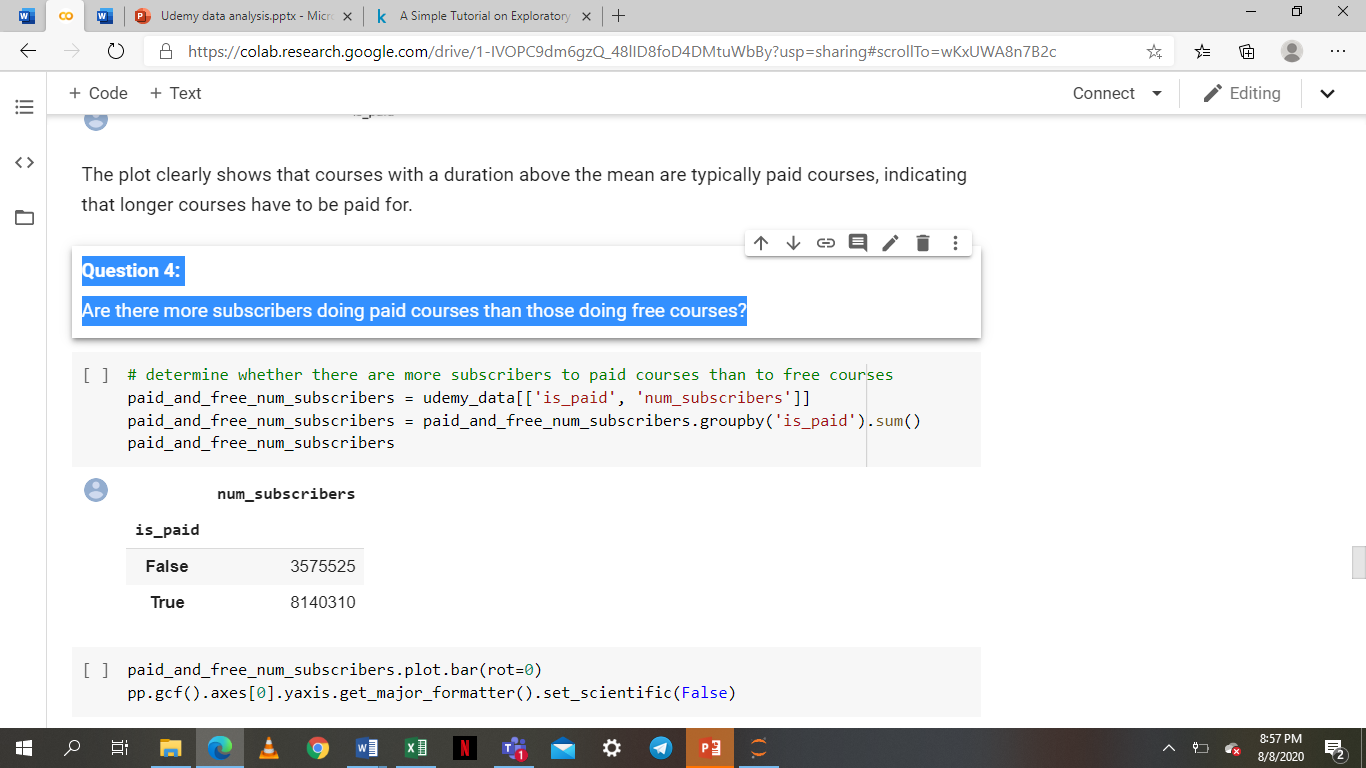


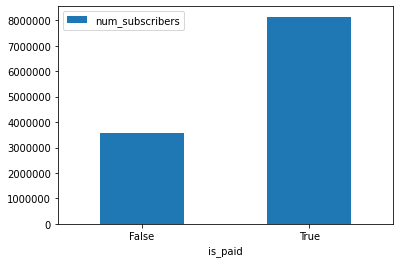


The plot clearly shows that courses with a duration above the mean are typically paid courses, indicating that longer courses have to be paid for.

**Question 4:**

Are there more subscribers doing paid courses than those doing free courses?

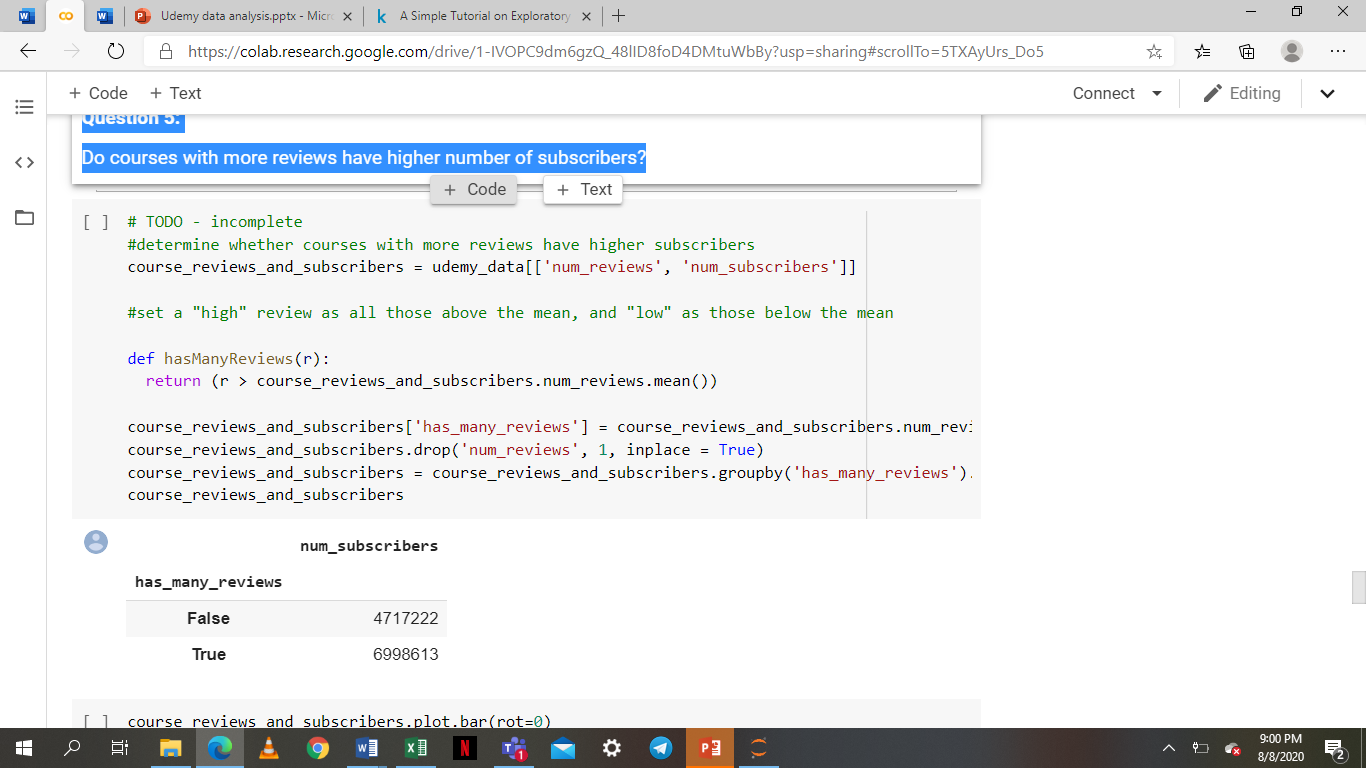


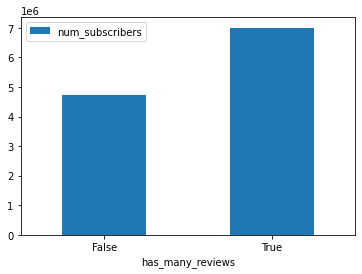


The plot shows that there are surprisingly more subscribers for paid courses than free courses.

**Question 5:**

Do courses with more reviews have higher number of subscribers?

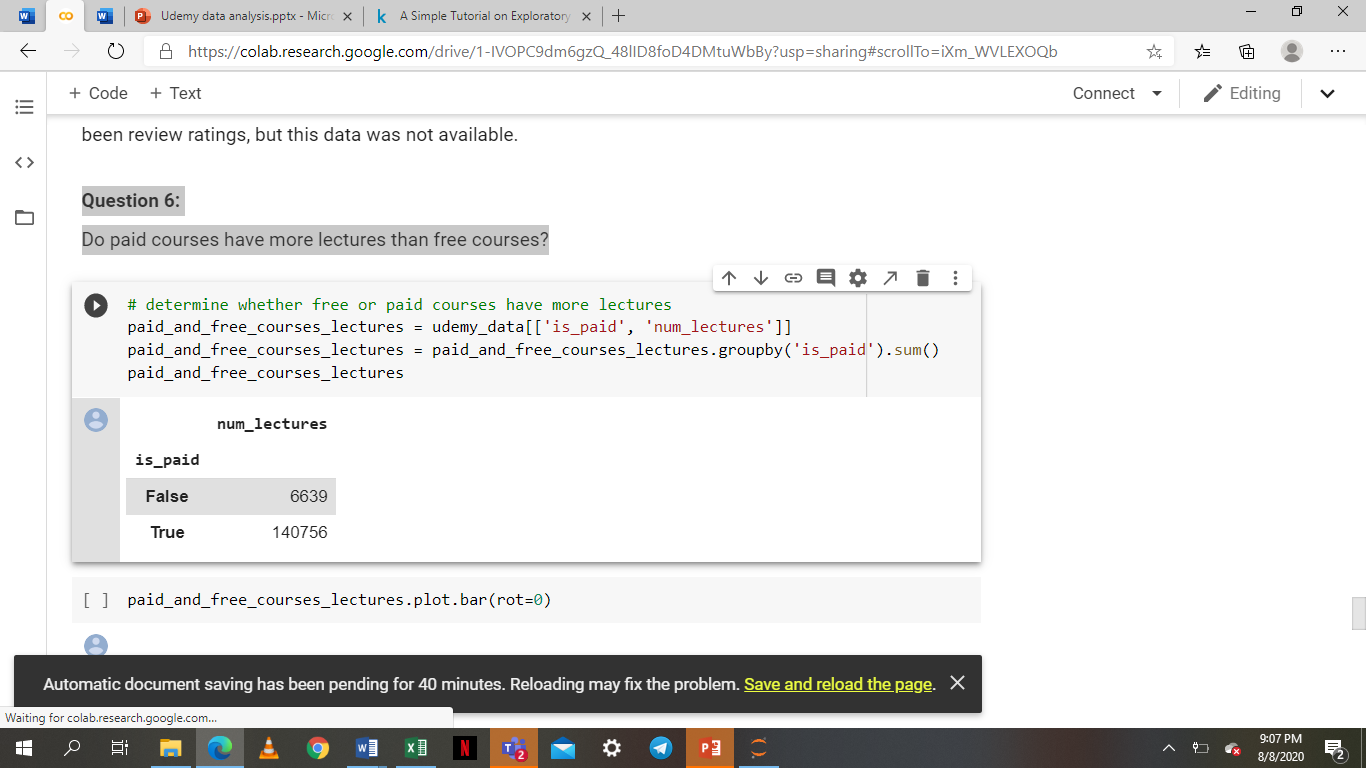


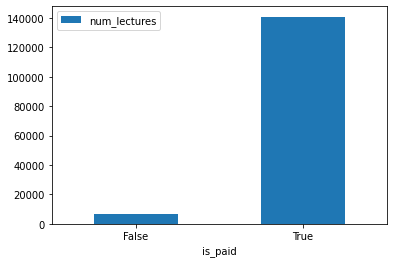


Intuitively, a course with higher number of subscribers should have more reviews. This is true from the graph. This cannot be used to gain much insight because more subscribers for a course would definitely mean the course gets more reviews. The factor which would have given more information would have been review ratings, but this data was not available.

**Question 6:**

Do paid courses have more lectures than free courses?

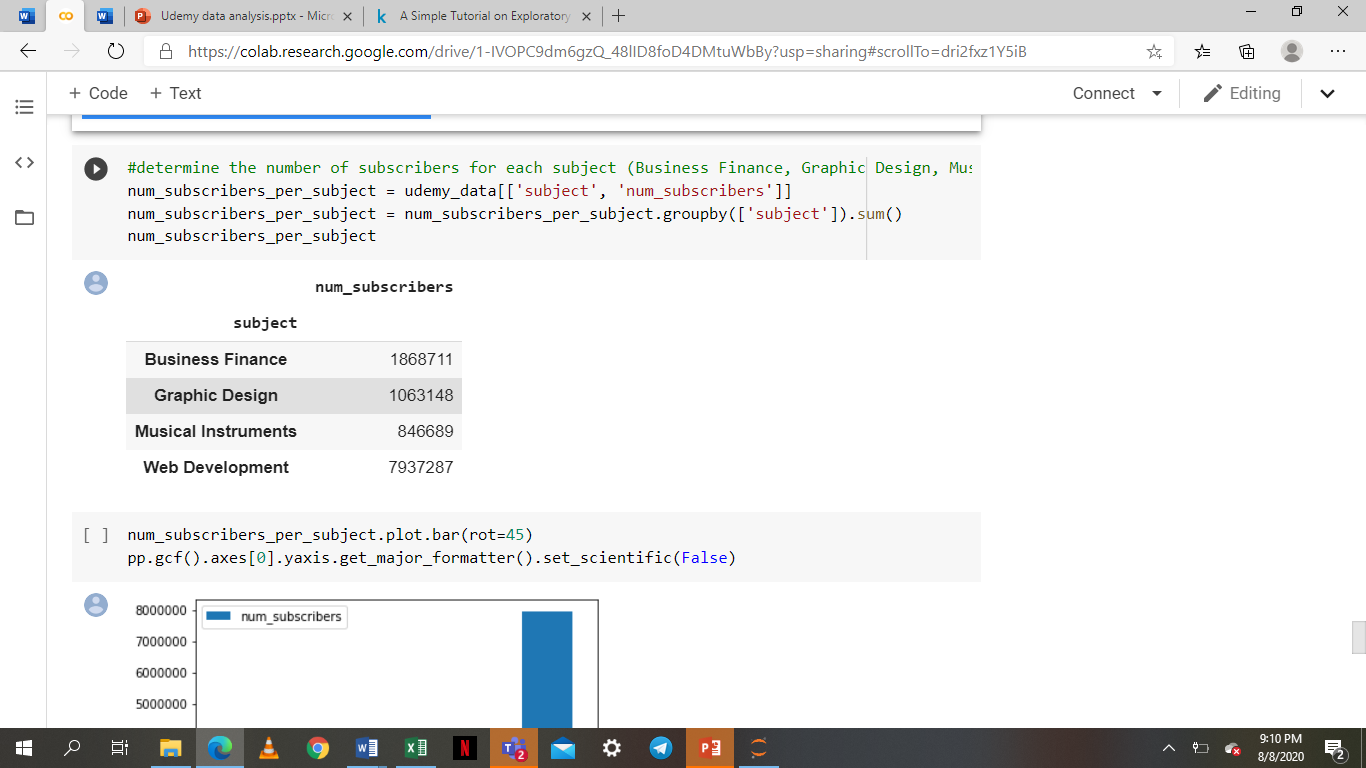


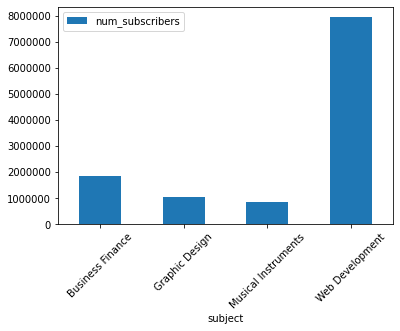


Paid courses are seen to have far more lectures in each course than the free ones, which would make sense from a business point of view - more lectures should be paid for.

**Question 7:**

Which subject has the most subscribers?

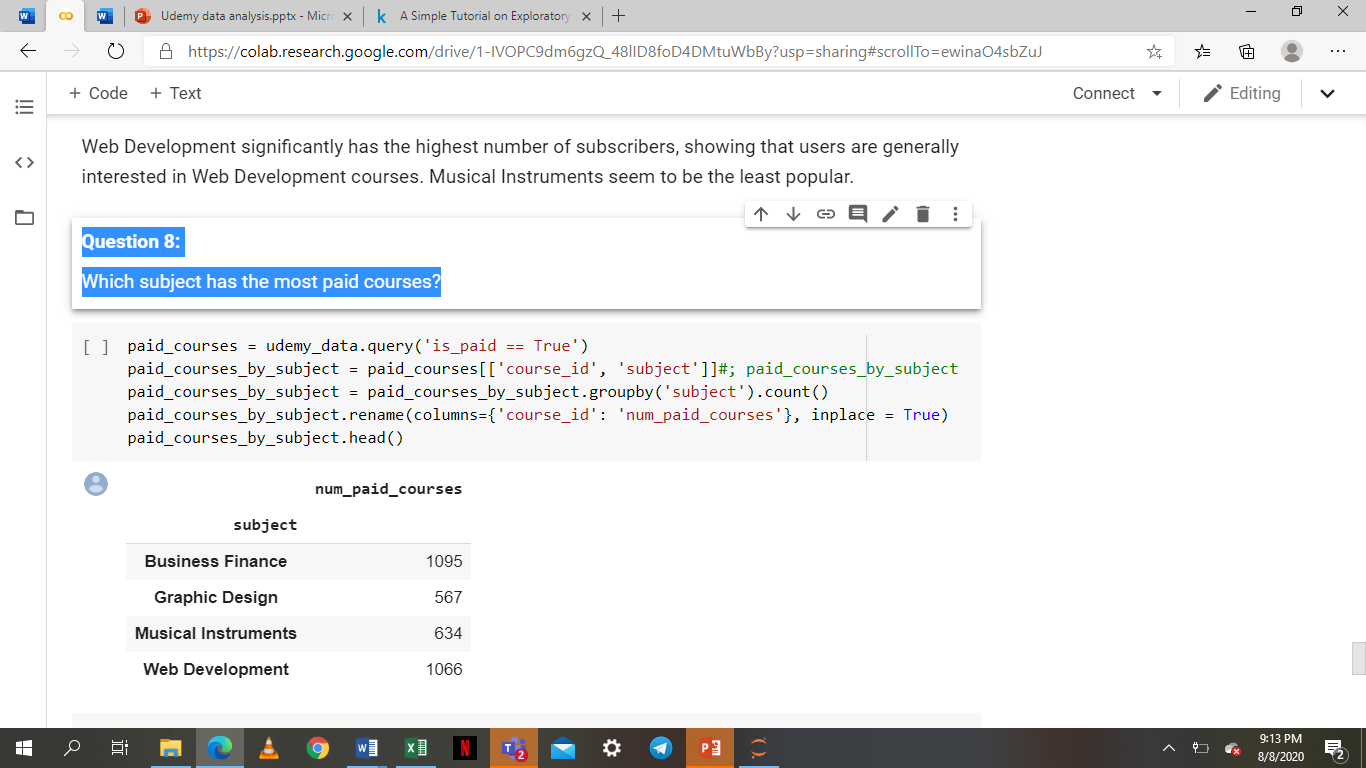


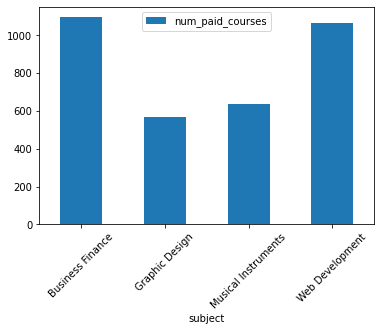


Web Development significantly has the highest number of subscribers, showing that users are generally interested in Web Development courses. Musical Instruments seem to be the least popular.

**Question 8:**

Which subject has the most paid courses?

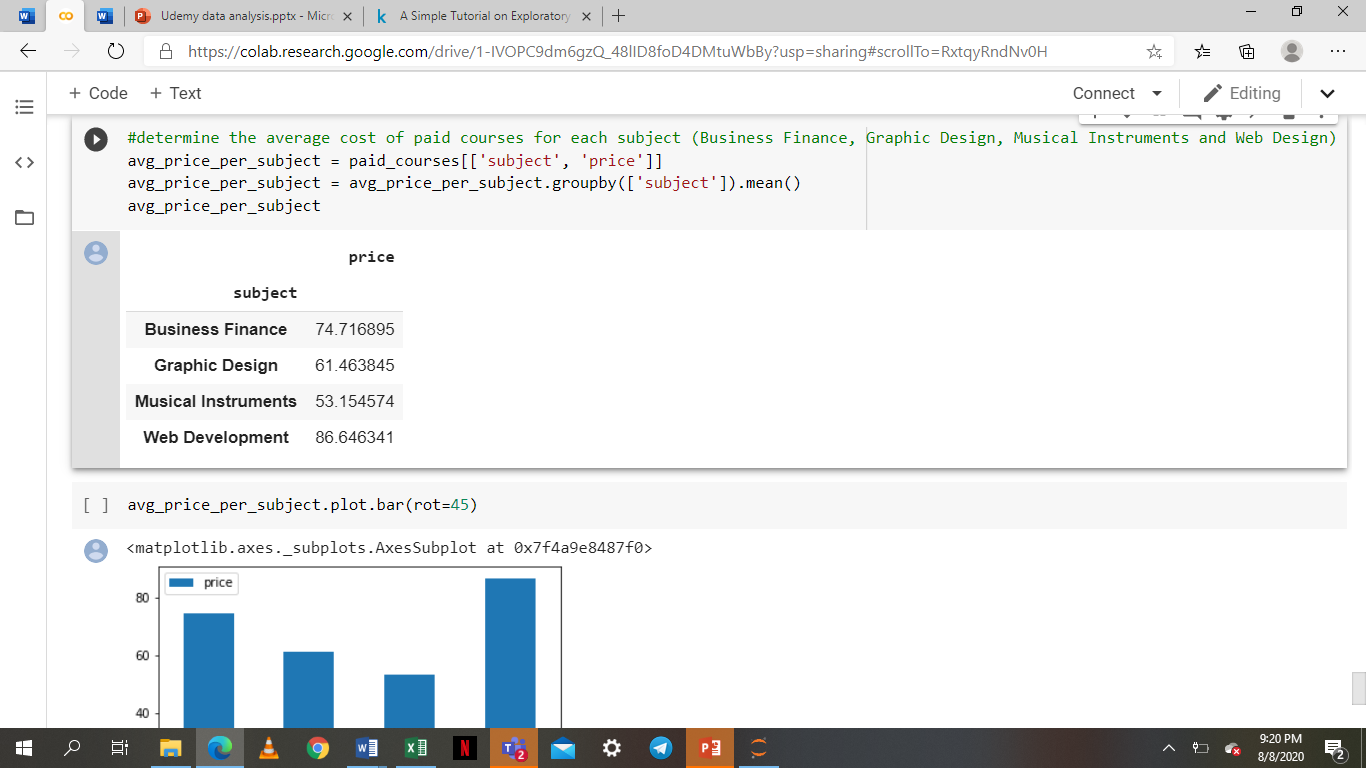


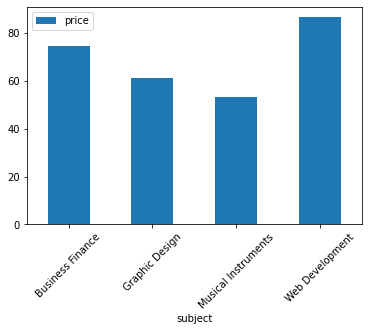


Business Finance subjects have the most paid courses, followed by Web Development. This result is interesting because Business Finance courses have less than half the number of subscribers that Web Development courses have. This may mean that users generally avoid Business Finance courses because they have to pay for them, although this cannot be proven without further analysis.

**Question 9:**

What were the average price for each subject?





Web Development courses were found to be averagely the most expensive, with Musical Instruments courses costing the least.

**FINDINGS**

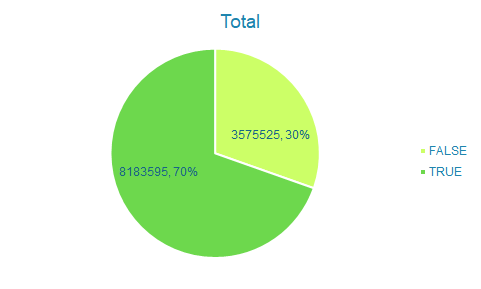
1. Findings from the number of subscribers for each subject

* Web Development significantly has the highest number of subscribers showing that users are generally interested in Web Development courses.
* Web development was the first subject introduced in 2011. In view of this, it has more subscribers.
* Musical Instruments seem to be the least popular.

1. From the given data, the number of subscribers increased from 2012 to 2015 and decreased in 2016 and 2017.The decreasing values are as a result of :

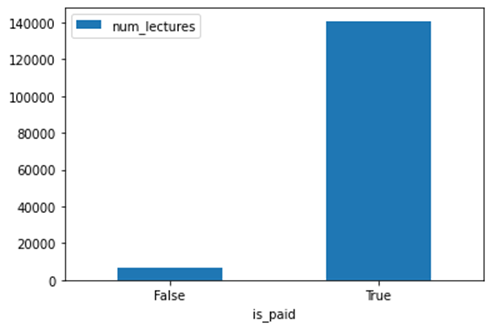
* New online learning platforms created offering the same services.
* Difficulty level of the courses are high i.e. for expert levels.
* Number of course reviews keeps decreasing.
* Duration of the course videos being long.

1. Number of subscribers for paid and free courses



From the pie chart, the number of subscribers for paid courses is 70% and that of free courses is 30%. The plot indicates that the number of paid subscribers is decreasing over the years. This may be due to the fact that overall number of subscribers are decreasing. However, it is noticeable that the number of subscribers taking free courses is far lower per year than the free ones. The number of subscribers taking paid courses is 40% more than those taking free courses.

1. Number of lectures for free and paid courses



From the graph, paid courses have far more lectures in each course than the free ones. This shows that business is making progress.

1. Findings from the level of difficulty of courses

* Expert level courses significantly have the least subscribers.
* Intermediate level courses also have less than half the number of subscribers that beginner level courses have.
* This indicates that more people subscribe to beginner level courses.

1. Findings from subject with the most paid courses and average price per subject

* Business Finance subjects have the most paid courses yet have less than half the number of subscribers that Web Development courses have.
* Web Development courses have the most number of subscribers and are averagely the most expensive.

**CHALLENGES**

* Difficulty in differentiating between hypothesis and questions.
* Difficulty in defining the problem from the dataset provided.

**HOW THE CHALLENGES WERE SOLVED**

We researched more and asked our TeachOps for further explanation about hypothesis and questions. With a better understanding into the hypothesis and questions, we were able to find the problem. An understanding of the business and finding the independent and dependent variables helped us find the main problem.

**LESSONS FROM THE CHALLENGES**

* It is good a thing to ask for clarification when given problem to solve.
* When given a dataset to work on you need to know what is expected of you and you have work hard towards achieving the goal.