

**A Project Report On**

**Udemy Courses Data Analysis using Python**

**By**

**Rajashri G. Tapkir**

**BE(Computer Science)**

**Batch : 2021 – 5127**

**Center: Pune**

**Under the Guidance of,**

**Mr. Mathivanan Balakrishnan**

**Technical Trainer**

**EduBridge**

(School of coding)

**Contents:**

* Software Requrements
* Introduction
* Data Preparation & Cleaning
* Data Analysis & Visualization

1. What are the udemy subjects udemy offers?
2. Most Popular Courses by num\_subscribers
3. Count of Paid and Free Courses?
4. Show all the subject : price vs subscribers by line graph
5. Price Count using bar graph
6. Majority price is in between price 25-60 using boxplot
7. Scatter plot of price by course level
8. Summary of the numerical attributes by histogram
9. Plot subject by number of subscribers and level
10. Plotting Heatmap
11. Linear Regression

* Conclusion

**Software Requirments:**

* Software: Anaconda Navigator , Jupyter
* Operating System : Windows 10 , 64-bit

**Introduction:**

Udemy, founded in May 2010, is an American online learning platform aimed at professional adults and students. As of Jan 2020, the platform has more than 50 million students and 57,000 instructors teaching courses in over 65 languages. There have been over 295 million course enrollments. Students and instructors come from 190+ countries and 2/3 of the students are located outside of the U.S.

Students take courses largely as a means of improving job-related skills. Some courses generate credit toward technical certification. Udemy has made a special effort to attract corporate trainers seeking to create coursework for employees of their company. As of 2020, there are more than 150,000 courses on the website.

Udemy dataset contains **3678 records** of courses from **4 subjects (Business Finance, Graphic Design, Musical Instruments and Web Developement)** taken from Udemy. Udemy is a massive online open course (MOOC) platform that offers both free and paid courses. Anybody can create a course, a business model by which allowed Udemy to have hundreds of thousands of courses.

**Udemy dataset for the year 2011-2017 helps to analyze & visualize which course is highly viewed or purchased, which courses udemy should focus on, when udemy should publish their new courses.** This version modifies column names, removes empty columns and aggregates everything into a single csv file for ease of use.

The project aims to convey the analysis of different courses is paid or not from the data from the internet. The python libraries used in the exploratory data analysis include NumPy, Pandas, Matplotlib, Seaborn.The objectives for the following project include:

**Data Preparation & Cleaning:**

We read the data file and aggregate the data on a few fields (course\_id, course\_title, url, is\_paid, price, num\_subscribers, num\_lectures, level, content\_duration, published\_timestamp, subject). Data Cleaning is the most crucial step towards a successful data analysis project. In most of the cases, the dataset has few “NaN”(not a number) values, some empty rows(having value 0) as well as redundant columns which could be removed using and configuring drop function and changing NaN values to 0 or removing the entire row as per need.

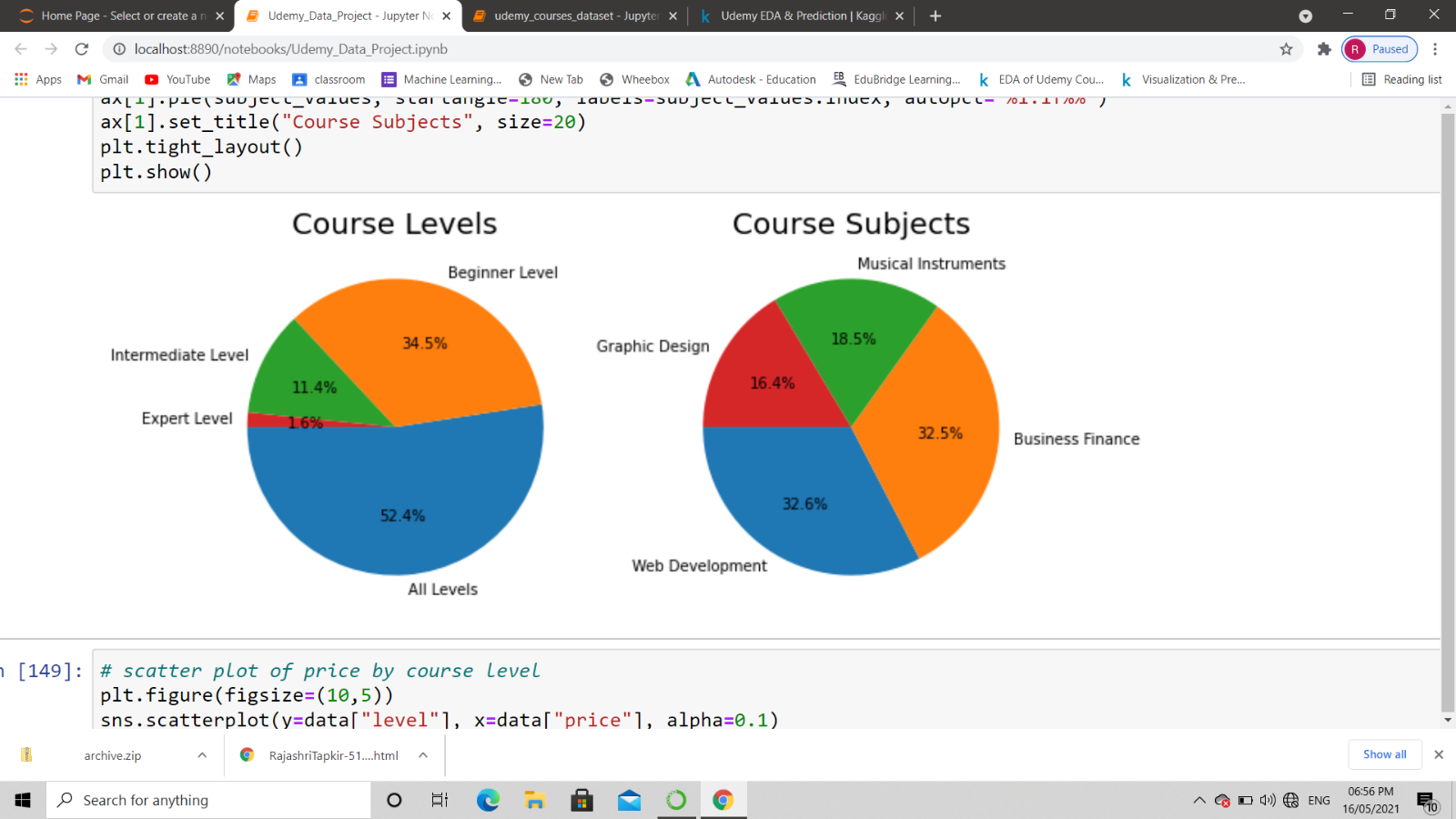
The reason for removing these columns is for the model to generalize better. Furthermore, these columns have a unique value for each instance (i.e. URL, course ID) which does not provide information the model can learn from to predict on new data.

* course\_id
* course\_title
* url
* published\_timestamp

**Data Analysis:**

**1.What are the udemy subjects udemy offers?**

Udemy offers the 4 subject there are Business Finance, Graphic Design, Musical Instruments and Web Development.

****

**2.****Which subject has Maximum number of courses each category has?**

The analysis of the subject has maximum number of courses each category has given below:

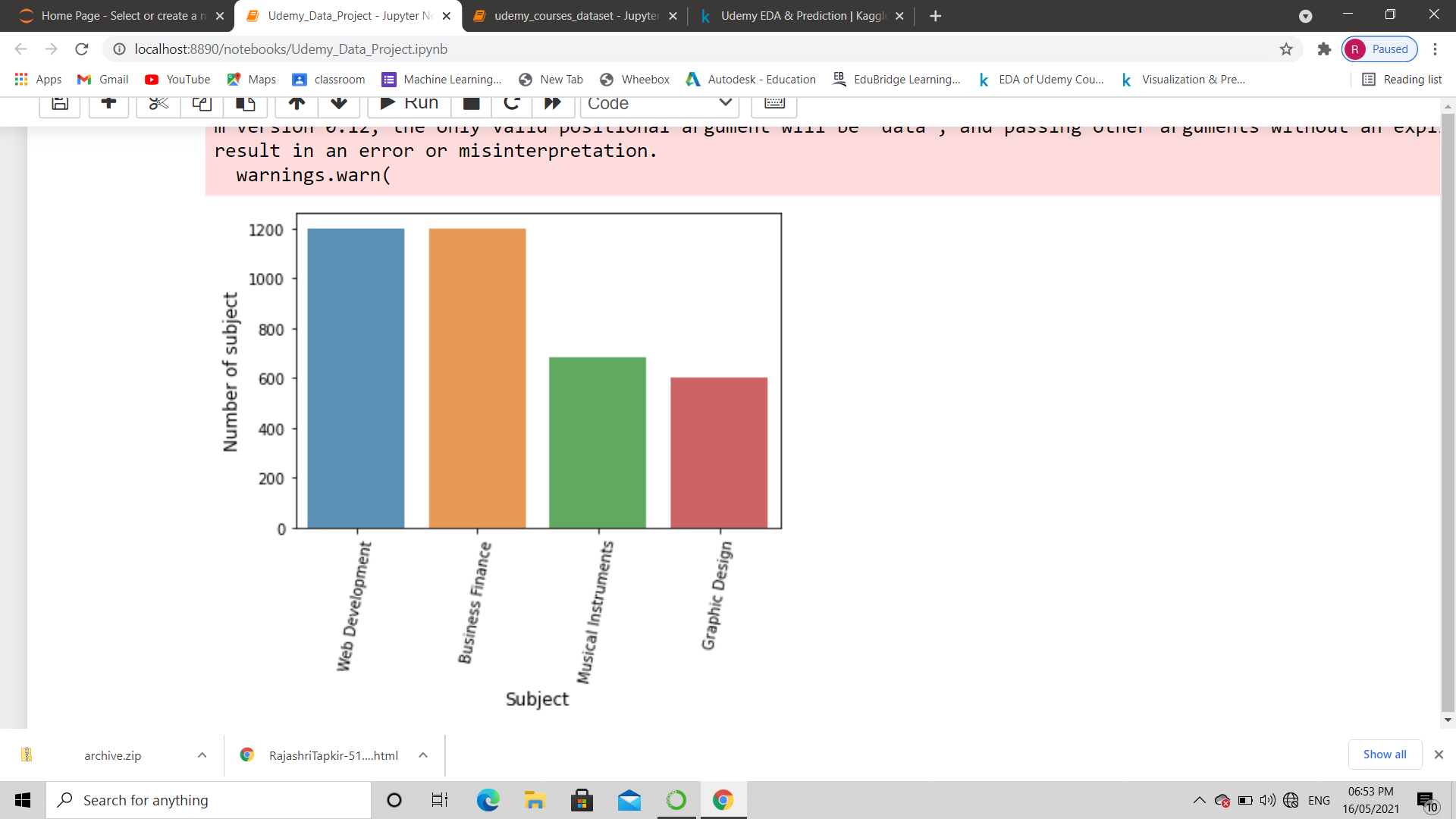
1. Web Development 1200

2. Business Finance 1195

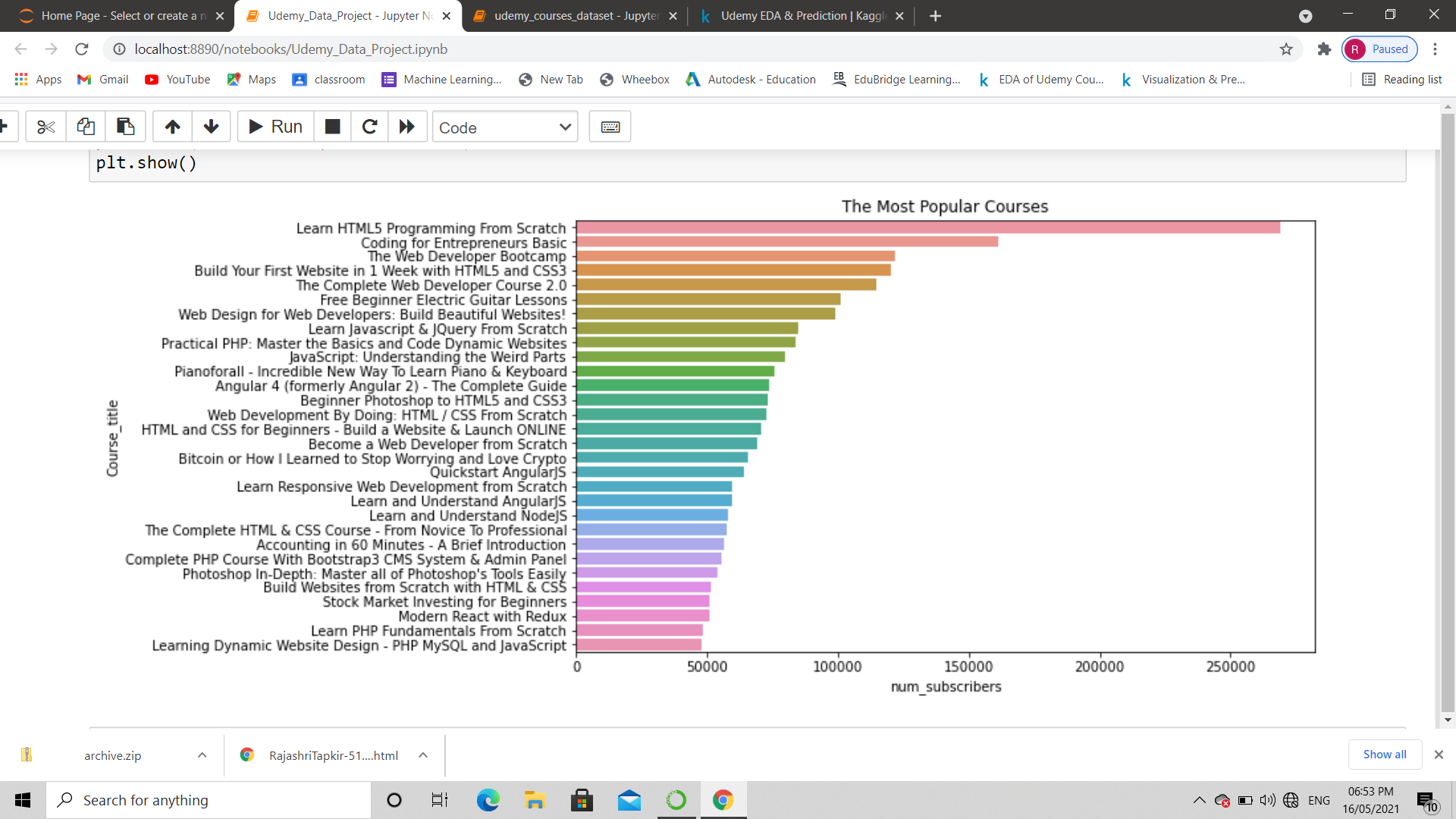
3. Musical Instruments 680

4. Graphic Design 603

As expected analysis ,web Development has maximum number of courses.

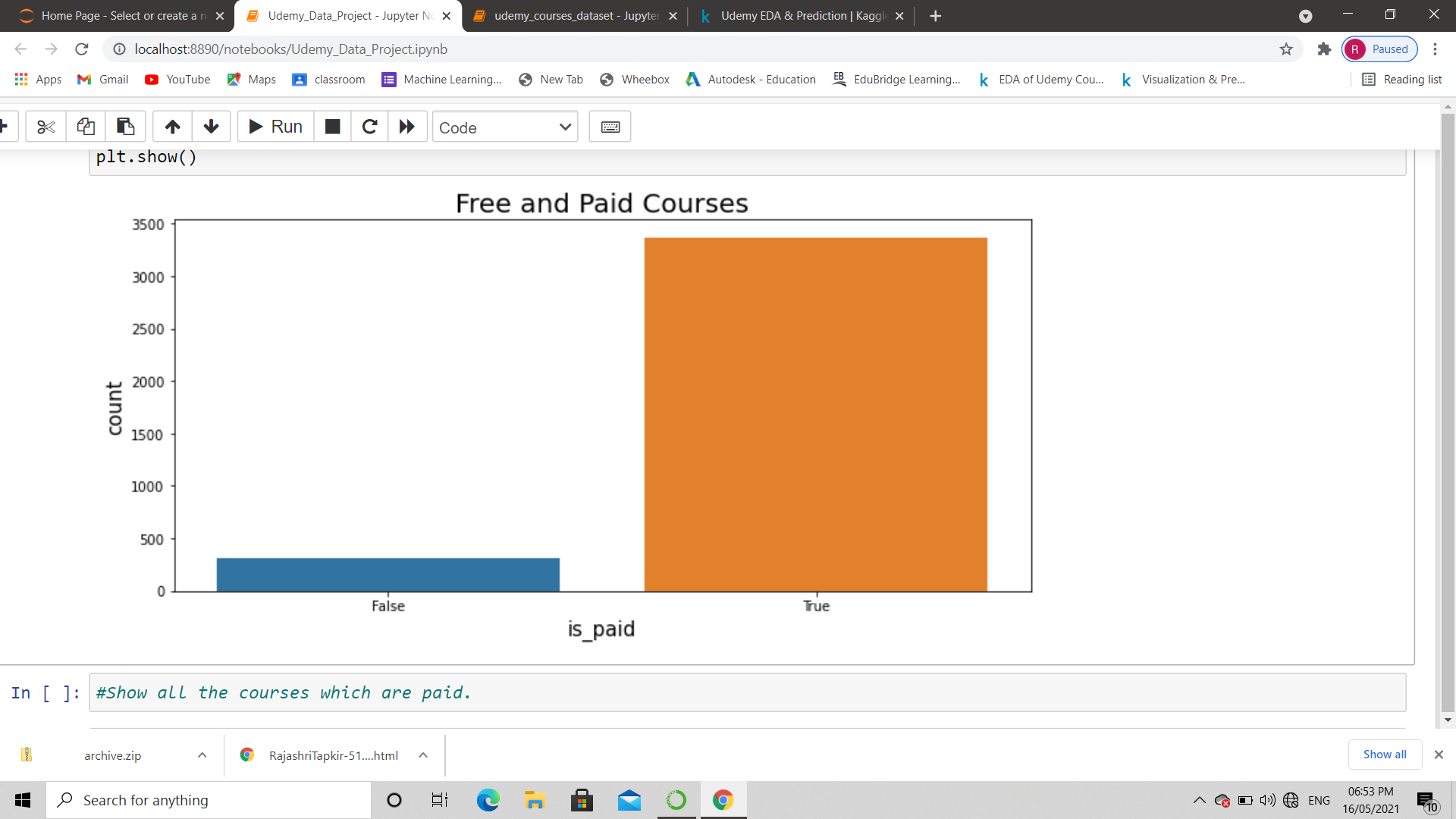
****

**3****.** **Most Popular Courses by num\_subscribers**

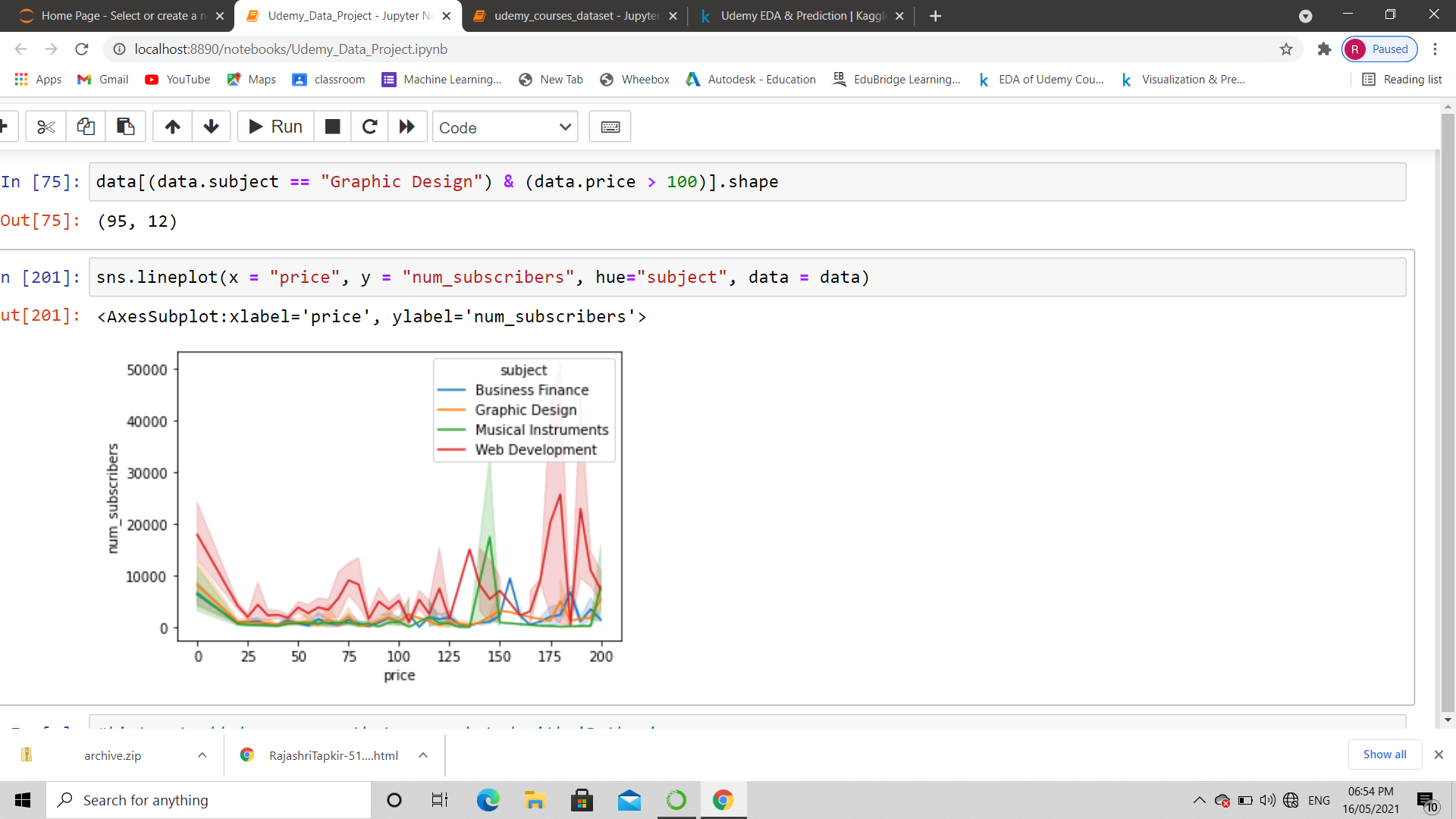
**** As expected analysis, the most popular courses is showing in the below Horizontal bar graph. The Learn HTML5 Programming from scratch is the most popular course from the analysis.

**4.****Count of Paid and Free Courses?**

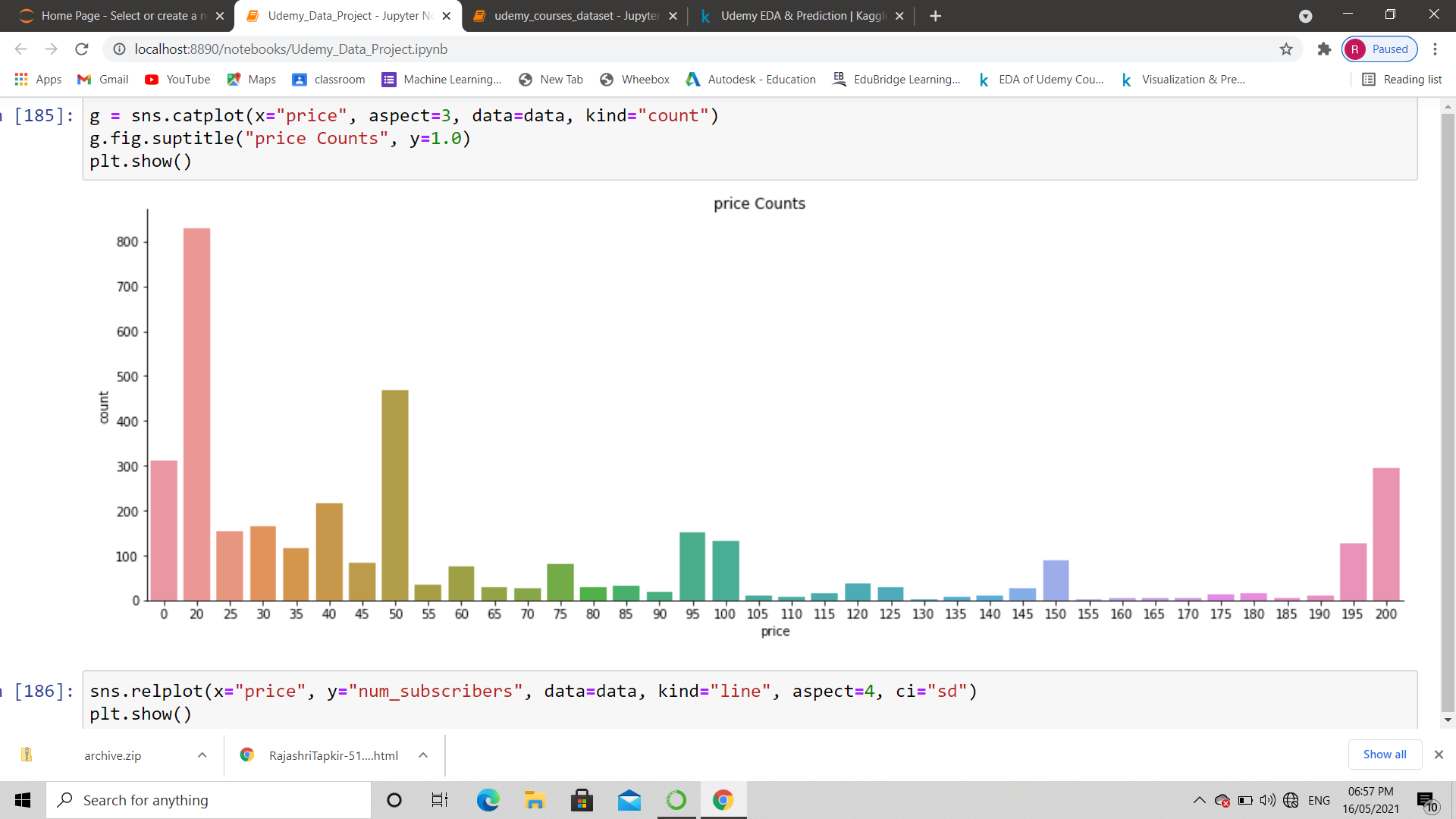
The False value means the course is free and the true value means the course is paid. The analysis of below graph paid courses are more than free courses. Paid courses count is larger than free courses.

****

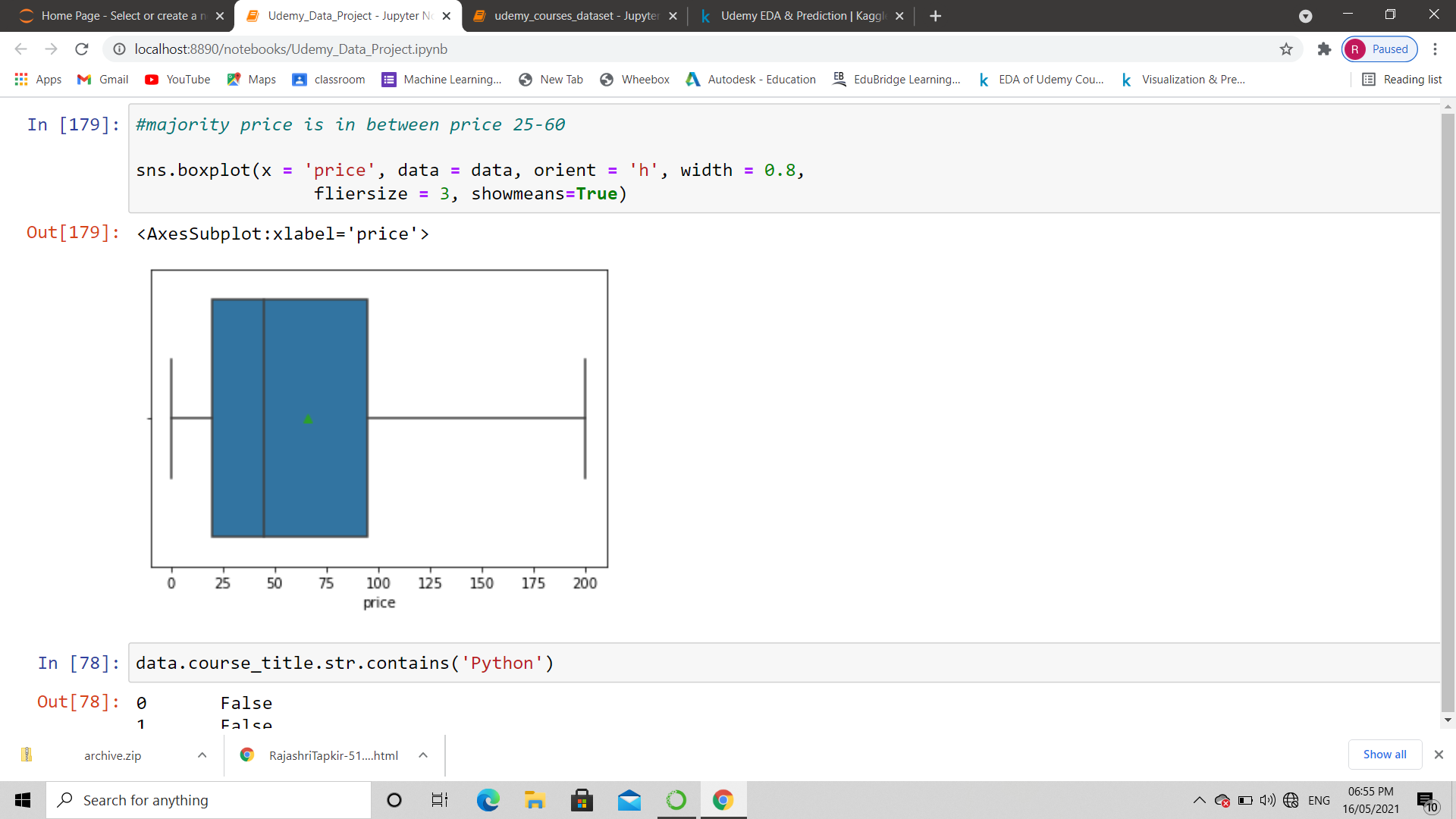
**5.****Show all the subject : price vs subscribers by line graph**

****

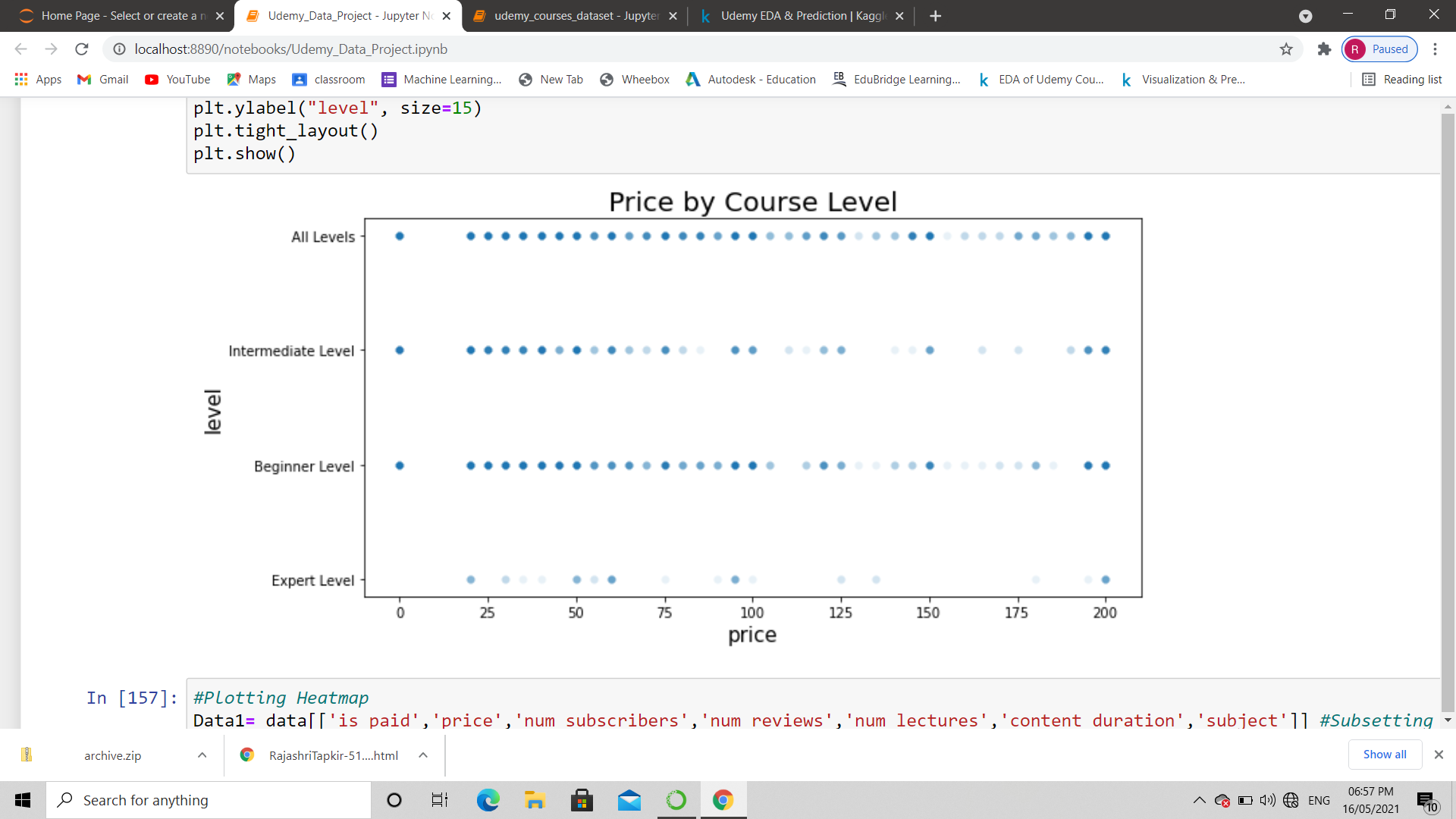
**6.****Price Count using bar graph**

****

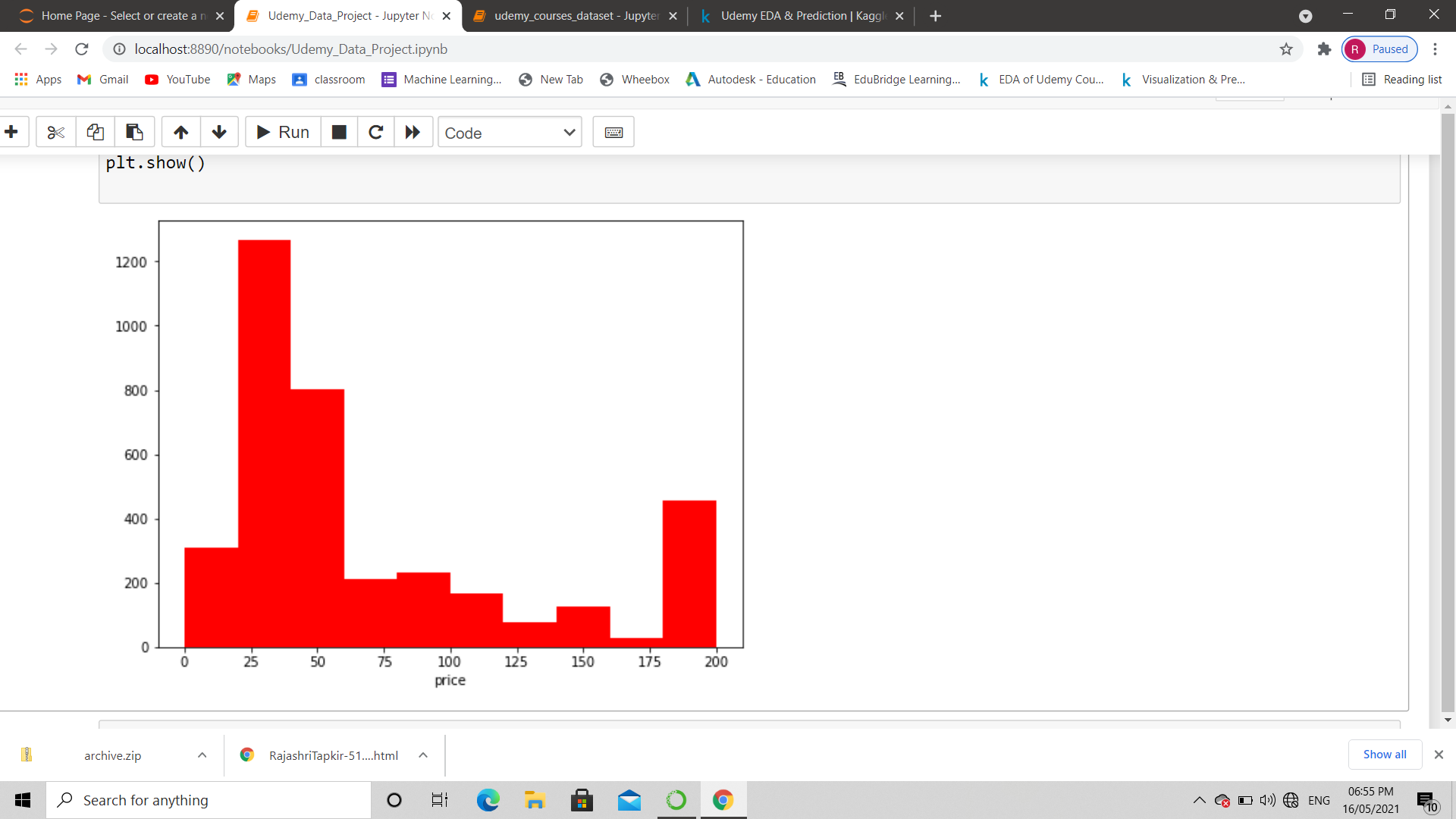
**7.** **Majority price is in between price 25-60 using boxplot**

****

**8. S****catter plot of price by course level**

****

**8.** **Summary of the numerical attributes by histogram**

****

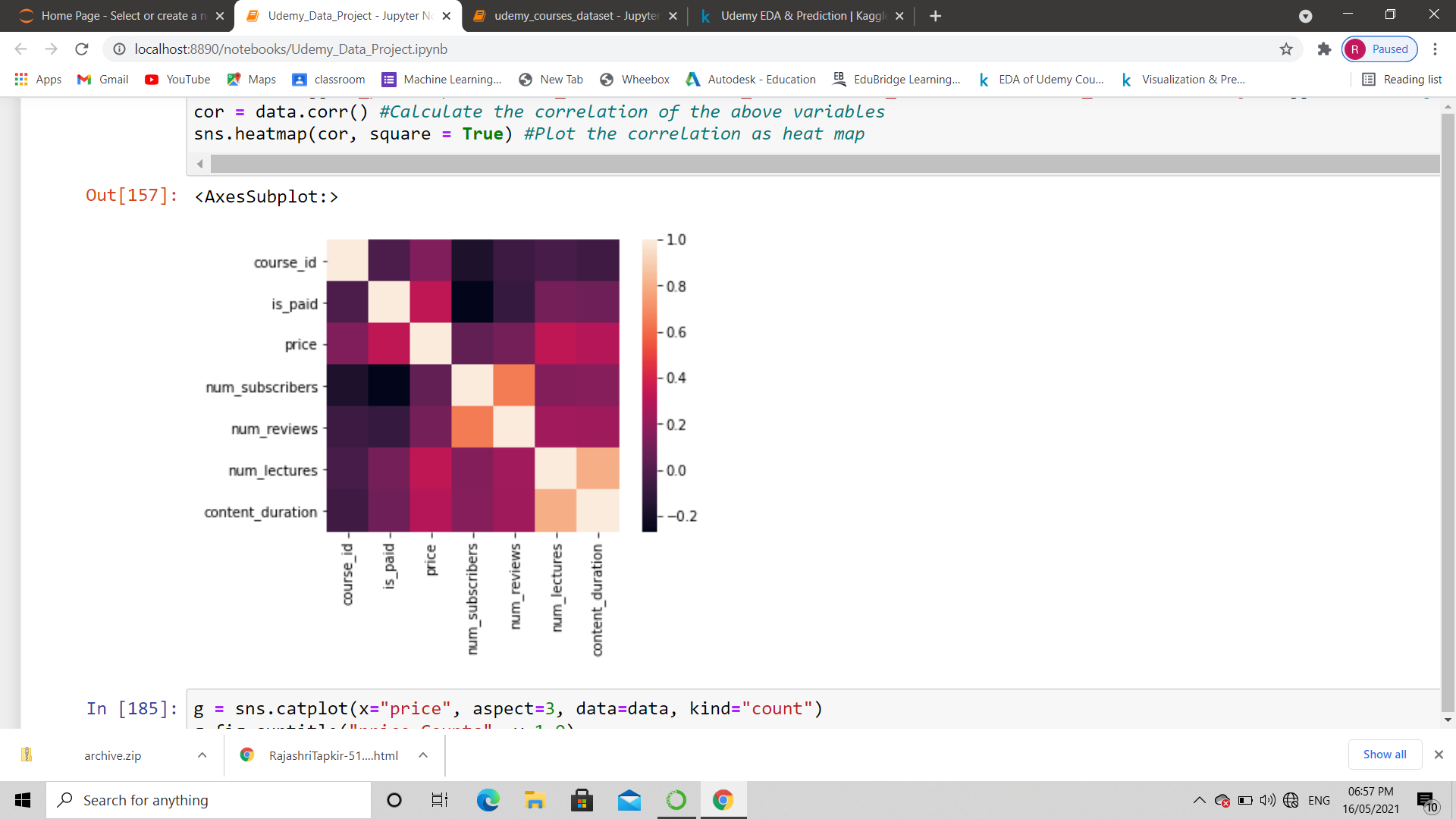
**9.****Plot subject by number of subscribers and level**

# 

* All Levels is the most common level, representing over 50%.
* Web Development is the most common subject, and Business Finance is second with approximately a 1% differential.
* Price variations according to the level of the course also show that Expert is the least common level in the data. It is also the only level that does not provide free courses. The other levels are dispersed more frequently throughout the line.
* Web Development courses are significantly higher in subscribers than the other subjects. Since Business Finance falls shortly behind in content, it is likely that people are more interested in studying Web Development courses.

# 10.Plotting Heatmap

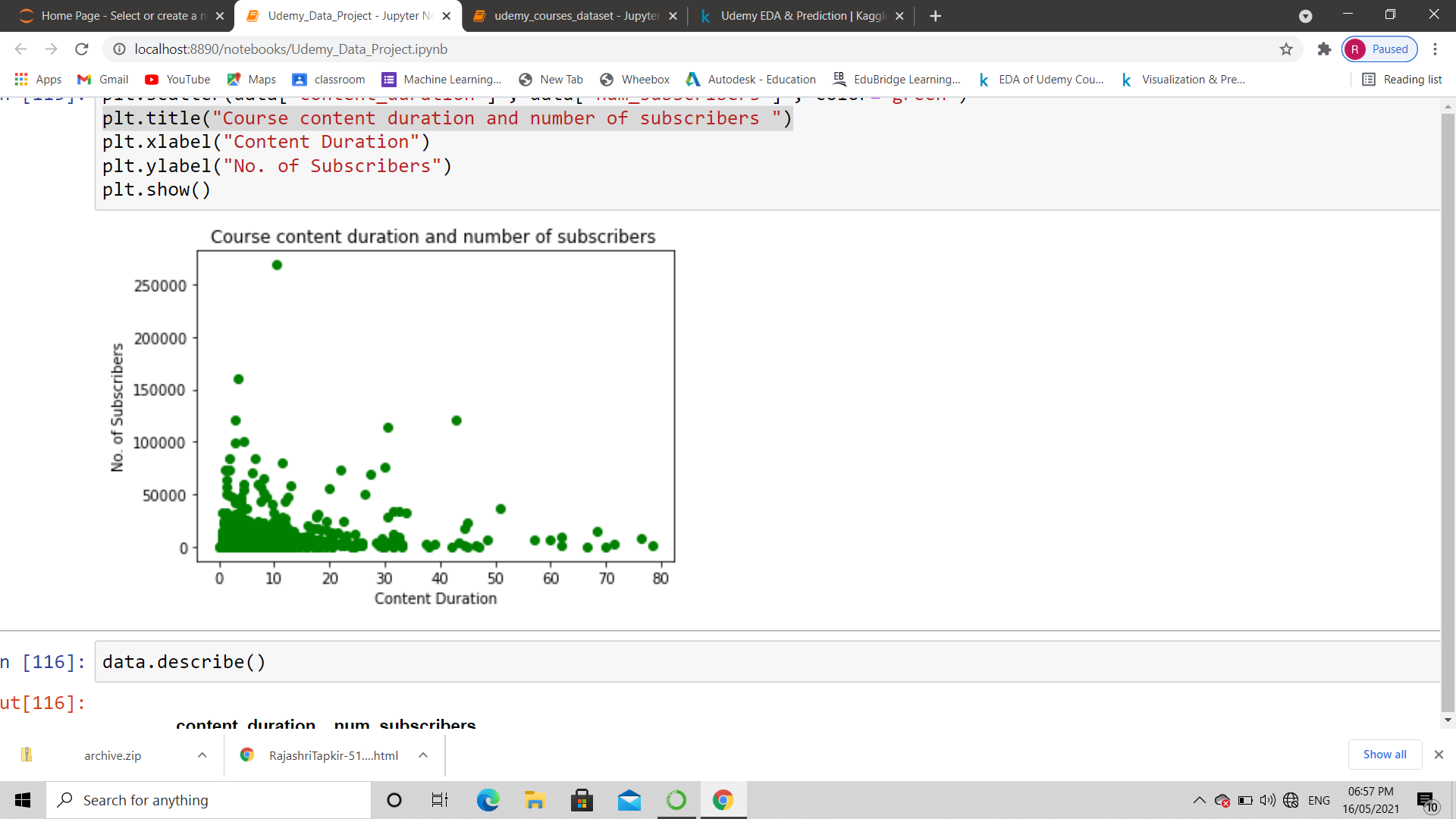
Heatmap can be defined as a method of graphically representing numerical data where individual data points contained in the matrix are represented using different colors. The colors in the heatmap can denote the frequency of an event, the performance of various metrics in the data set, and so on. Different color schemes are selected by varying businesses to present the data they want to be plotted on a heatmap.

****

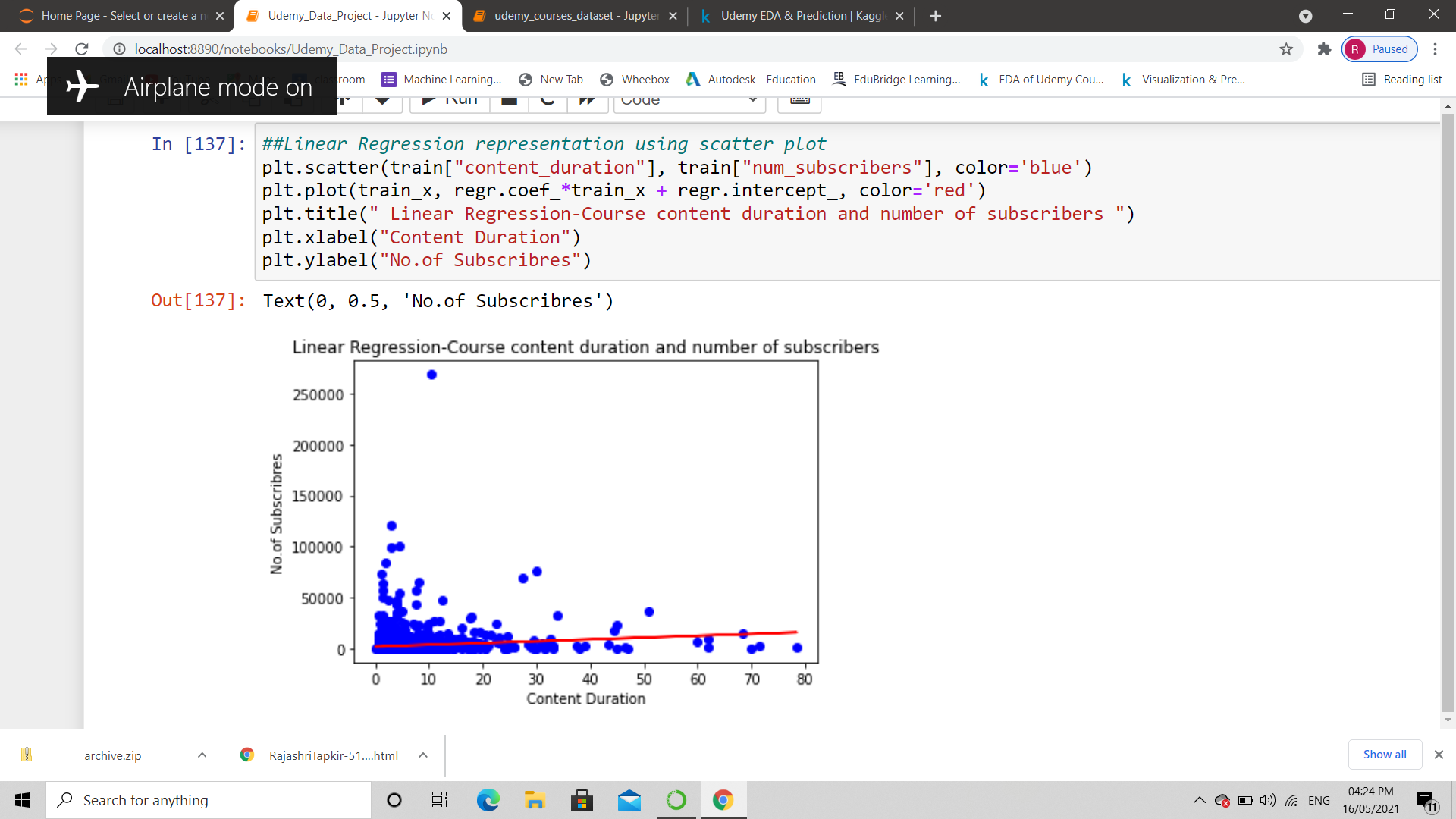
**11.****Linear Regression**

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

Scatter plot between course content duration and number of subscriber:

****

* Most of the usersprefer the courses to have less duration
* There is a linear relation between content duration and subscriber.

****

**Conclusion:**

We conclude that, 22.5% of all courses have a price of 20$, 12.7% of courses have 50$ price, 8% of courses are FREE, and also 8% of courses have 200$ price. The least of courses have the price more than 100$. Most of the courses have a duration of less than 10 hours. Mostly in the range between 1-5 hours.On average, every Udemy course has 25 lectures. Most of the courses are devoted to Web Development and Business Finance topics. However there are ore subscribers for Web Development and Musical Instruments. If we look at saperate levels, there are more courses for beginners. In general courses with All Levels and Beginner Level has the most subscribers. Most are the subcribers are subscribe for free courses. 1,3 million subscribers are of courses with 200$ and 20$. In general, the most subscribers are for FREE courses. The second place of popularity are courses with 200$. For Business Finance and Web Development some courses with the price 155 or 185 have also an increased number of subscribers. Web Development and Business Finance courses have the most lectures in total and the longest content duration. Most correlation is observed between Number of subscribers and Content Duration. The course for Web Development "Learn HTML 5 Programming from Scratch " is one of the popular on Udemy.