In this project, which I have made because of curiosity, I have investigated Udemy online courses platform data set. Due to COVID-19 lots of people started to learn something new or improving their skills online, so it is interesting to know what the most popular online learning platform can offer.

Unfortunatelly, the data I have found covers only 2011 — 2017 period of time, which is still fine to find out some patterns.

You can find the whole code in the folder and here are some insights I have found and also some of my assumptions.

Let's begin!

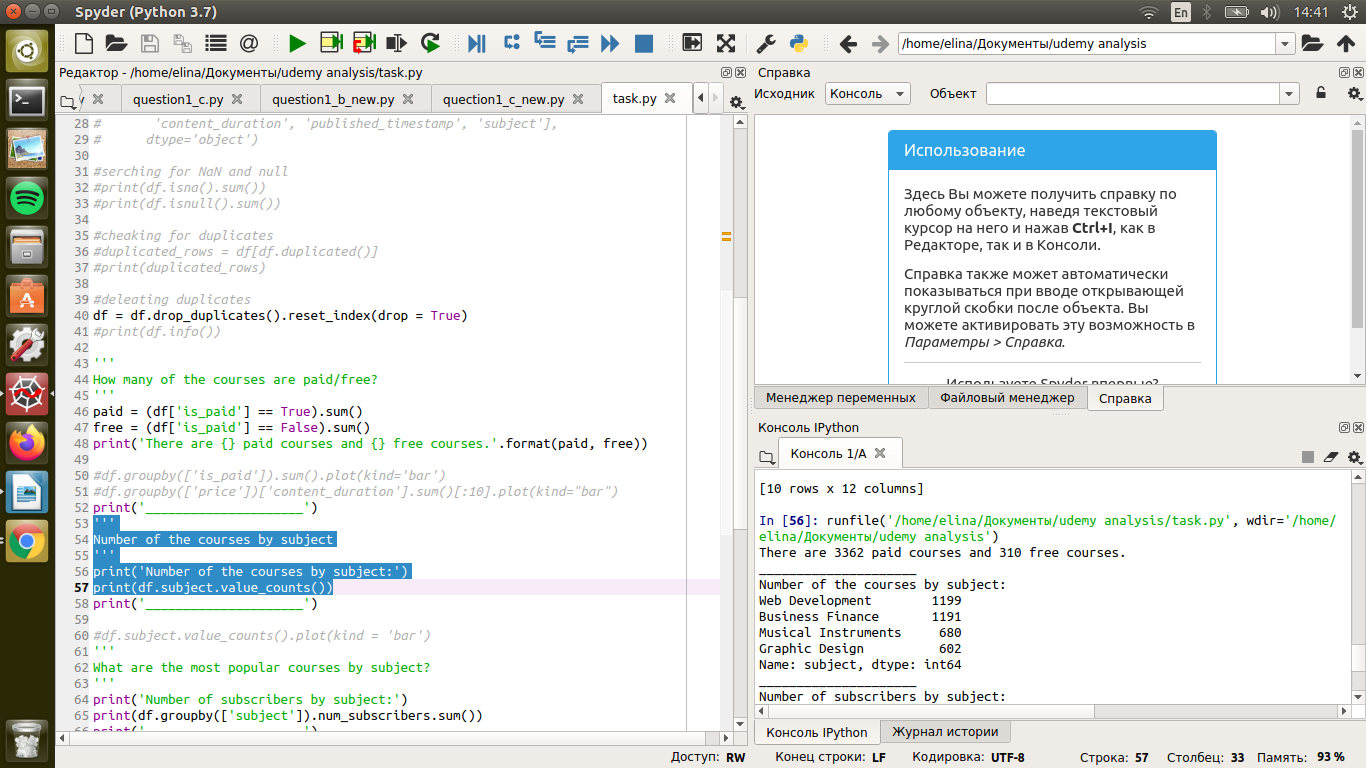
After checking the data for NaN values and duplicates we can see that it contains information on **3672 different courses.** The data show us courses in **four subject areas**:

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

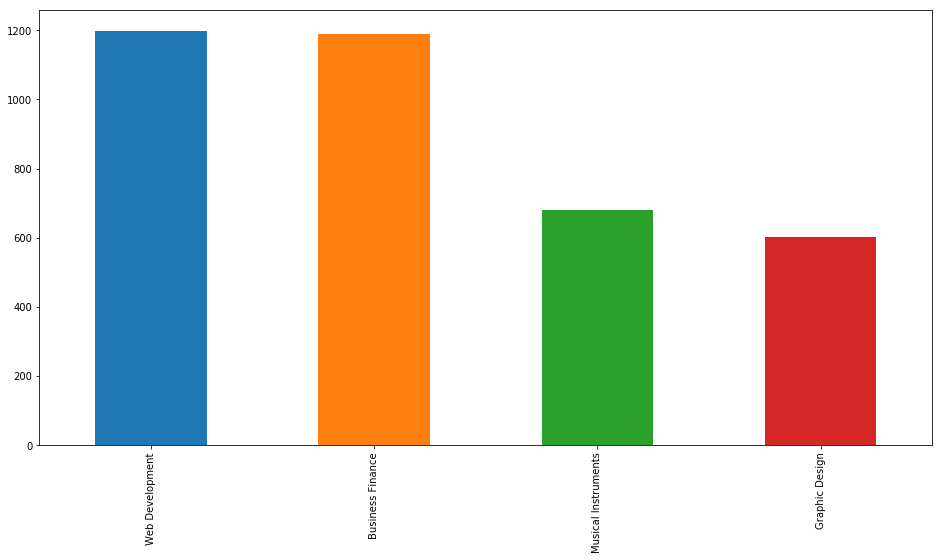
First of all, let's look at the number of courses by subject.

***print('Number of the courses by subject:')***

***print(df.subject.value\_counts())***



The visualisation always help to understand the results better.



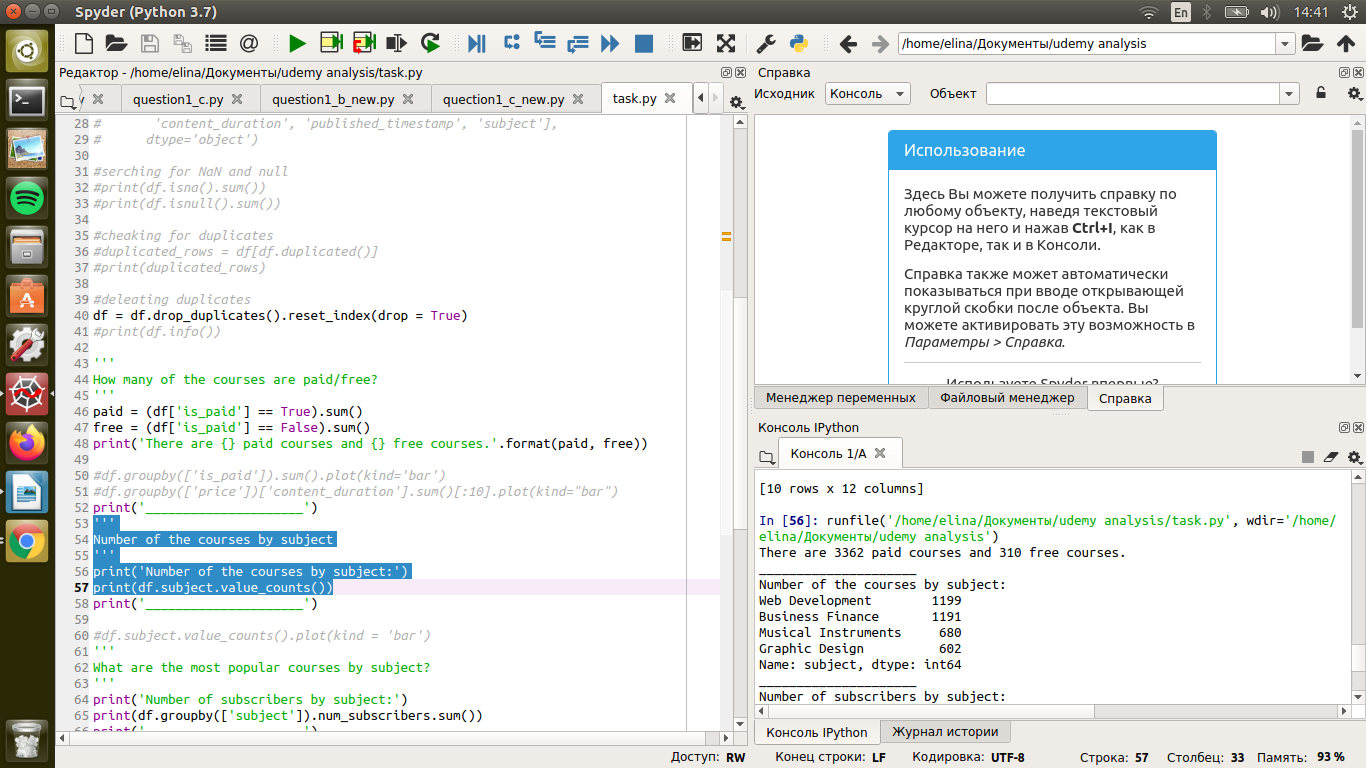
As we can see, the biggest part of the courses are **Web Developmen**t courses, nevertheless **Buisiness** **Finance** area has almost the same number of courses.

Based on the data there are paid and free courses. Let's look what Udemy can offer.

***paid = (df['is\_paid'] == True).sum()***

***free = (df['is\_paid'] == False).sum()***

***print('There are {} paid courses and {} free courses.'.format(paid, free))***

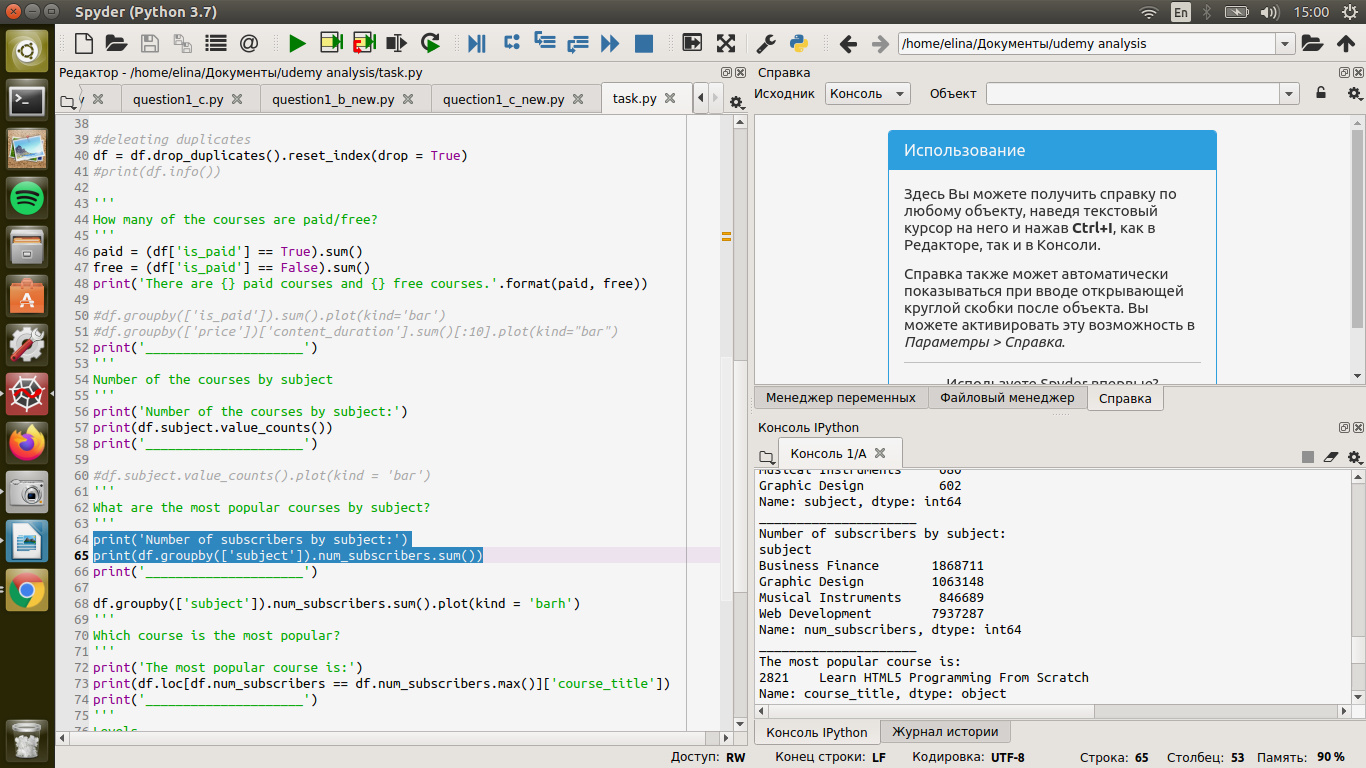
******

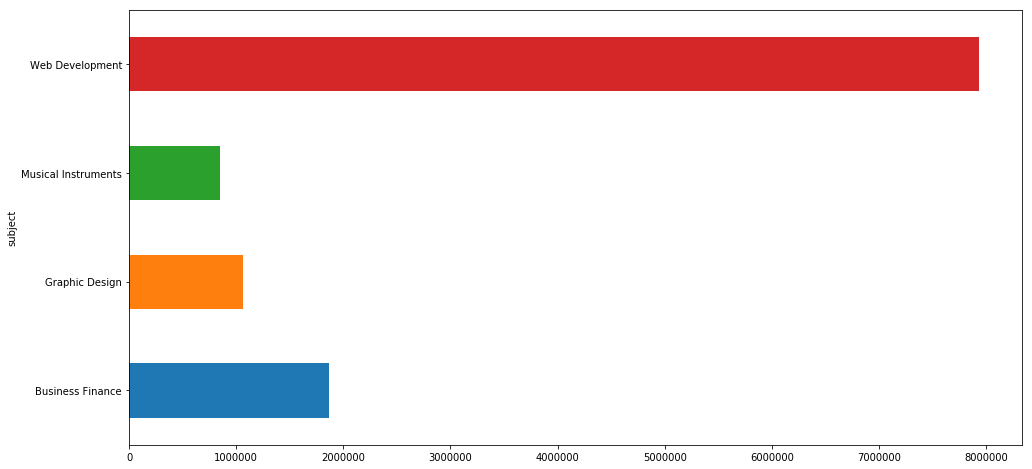
Well, we can get for free not that much.

Next, we will see what subject is the most popular for the Udemy's users. In other words, how many subscribers are there in every area.

***print('Number of subscribers by subject:')***

***print(df.groupby(['subject']).num\_subscribers.sum())***

******

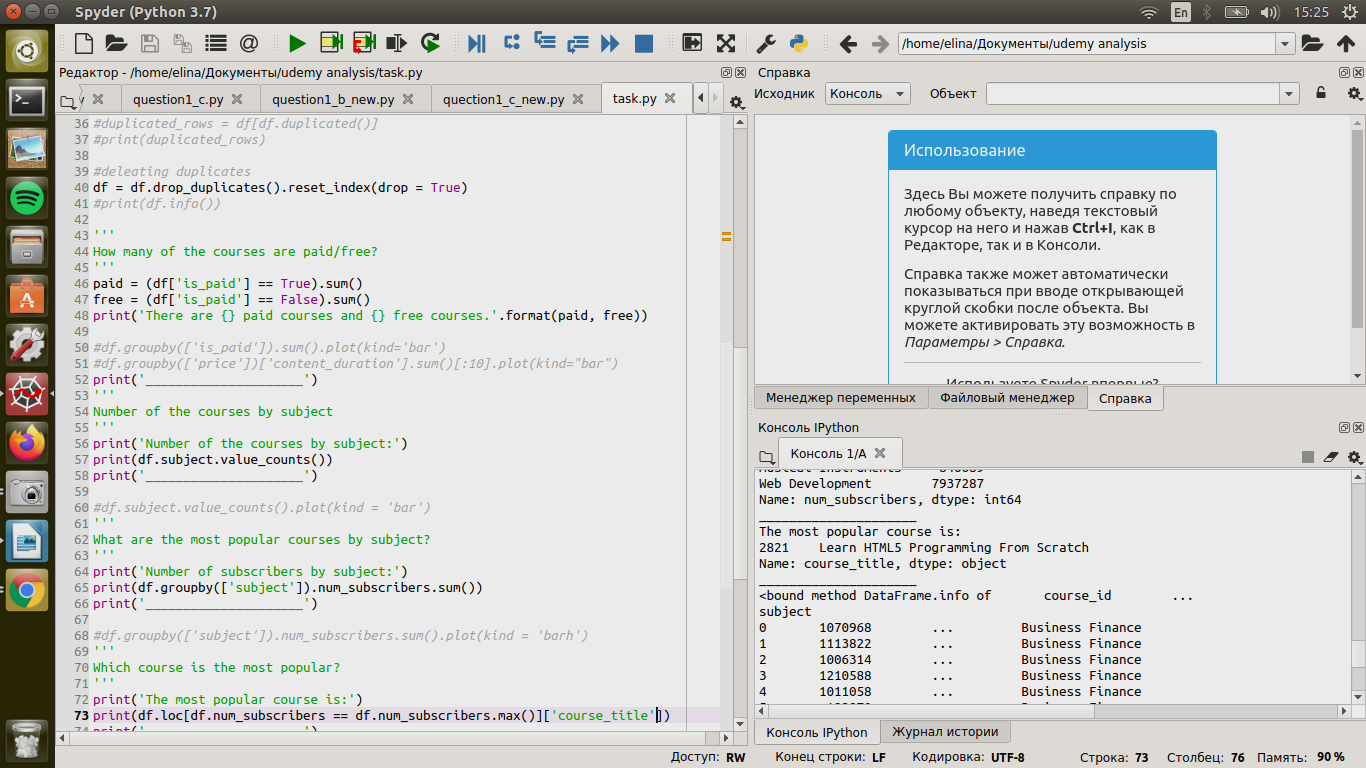


It is really easy to see that the **Web Development wins** with a large margin.

What is the most popular course? How many users are engeged to that course?

***print('The most popular course is:')***

***print(df.loc[df.num\_subscribers == df.num\_subscribers.max()]['course\_title'])***



All the courses have levels that are needed to start learning. Sometimes you need to already know some basic things or even be an expert in the particular field to take the course.

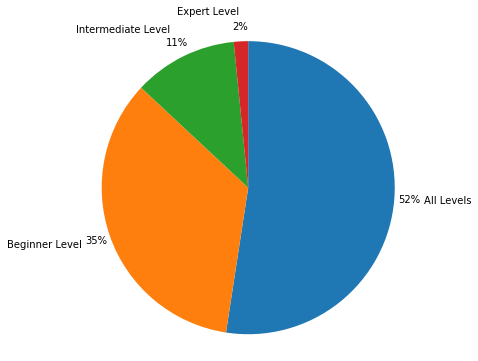
***plt.figure(figsize=(10,6))***

***sorted\_levels = df['level'].value\_counts()***

***plt.pie(sorted\_levels, labels = sorted\_levels.index, startangle = 90,***

***counterclock = False,autopct='%1.0f%%', pctdistance=1.1, labeldistance=1.2);***

***plt.axis('square')***



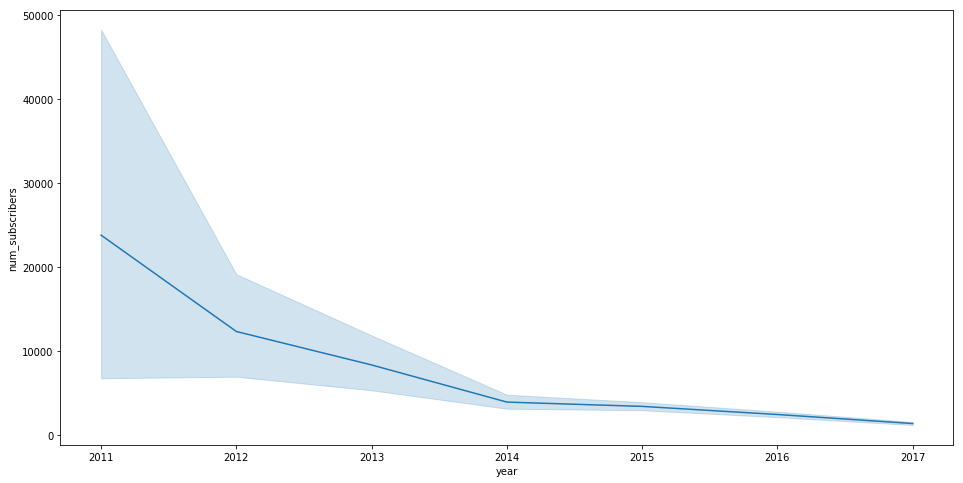
As we can see, more than a **half of the courses** are suited **for all levels** and **35%** are good **for** **beginners**. We can conclude, the platform suggests a wide range of skills if you want to learn something absolutely new.

Now it is time to find some correlations and patterns.

Let's find out how the number of subscribers was changing from year to year.

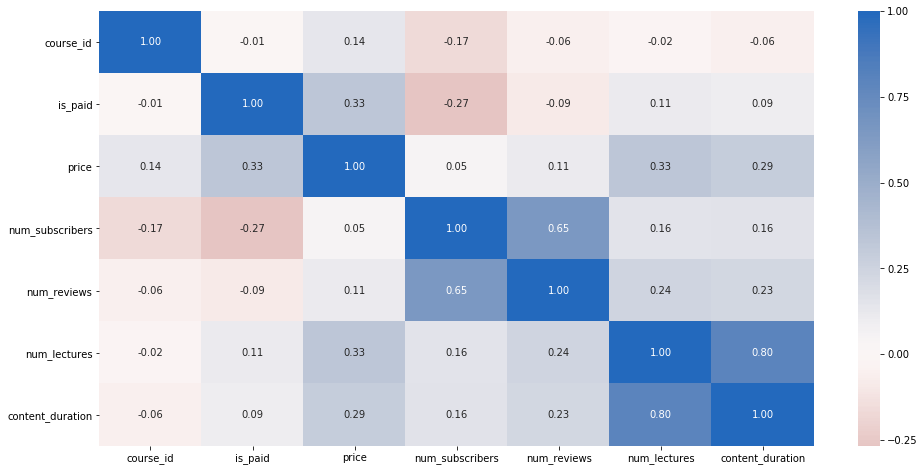
***df['year'] = pd.DatetimeIndex(df['published\_timestamp']).year***

***sns.lineplot(data = df, x = 'year', y = 'num\_subscribers')***

The number of subscribers **fell precipitously** (about 10000) from 2011 to 2012 and continued to fall up to 2014 and then it stays almost at **the same level**.

There is no data on the year 2020 but I suppose that the number of users has grown rapidly due to COVID-19 and quarantine.

Next, I would like to look at the way how all the columns are correlated to each other, in other words what of the parametres influence one another.

It is clear that **number of subscribers** and the fact is the course is **paid or free** is uncorrelated, which was confirmed by the chart (coefficient is -0.27).

Also **number of lectures** and just as **content duration** is correlated with the **price** of the course. So, the longer is the course, the higher is the price.

The **number of subscribers** is slightly correlated with the **price**, so we can say that **users prefer not** **really expensive courses** much more.

There were some insights I got from the Udemy Online Courses Platform data set. I believe that there is much more information I can get from this data deep down. So the project is going to be improved and expended.