

**TASK**

**Exploratory Data Analysis on the EdX\_2013 Academic Year Courses Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

The dataset shows users who were registered for different courses on Edx for the 2013 academic and this report will focus mainly on activity of users who actually got certified.

**DATA CLEANING**

While exploring the data set column by column checking unique values, visualizing counts for these unique values and inspecting row, we learn the following;

1. Columns ***last\_event\_DI*** to ***nforum\_posts*** show course activity, drop rows where all a NaNs. Without any course activity a user won’t be certified.
2. A number of columns are unnecessary for this EDA and will be deleted
3. Age column has special characters to be turned to NaNs and also must parsed to numeric.
4. 3 more columns must be added; ***age\_group,*** ***duration*** and ***duration\_range***.
5. No duplicate rows were found.

Functions used include;

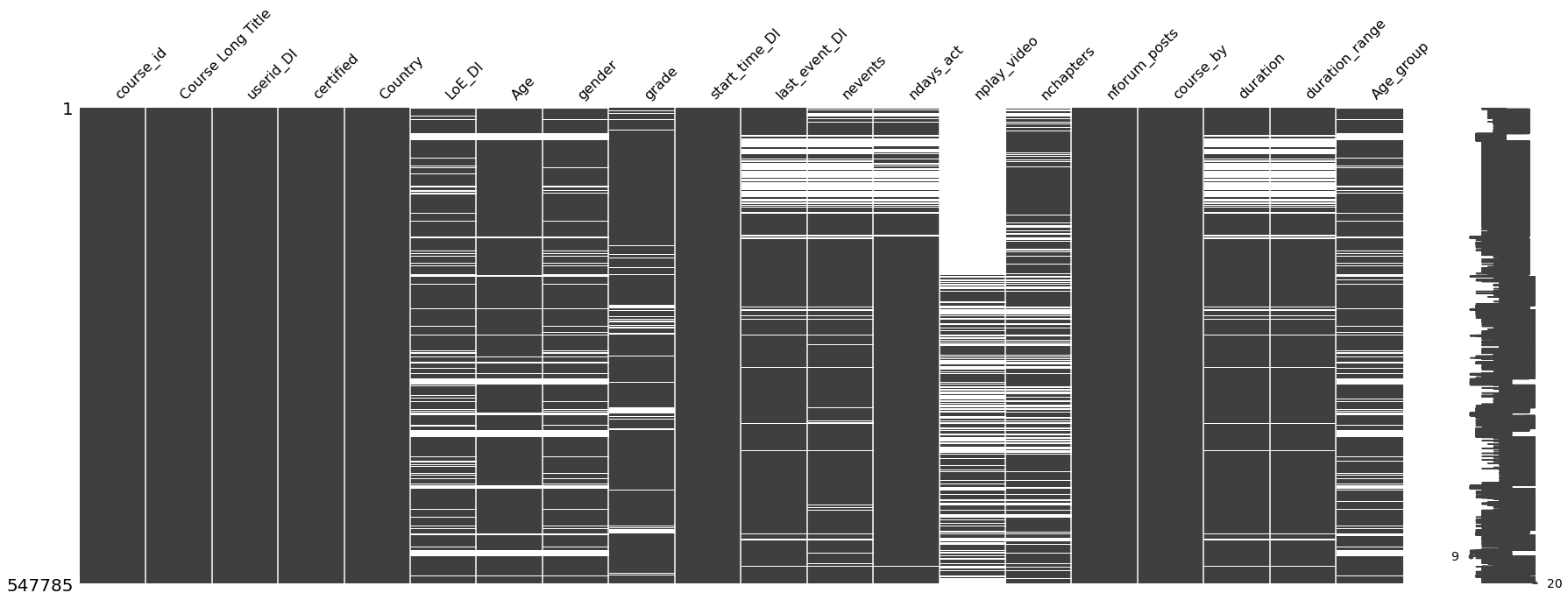
1. DataFrame.drop() for droping rows or columns.
2. DataFrame['column\_name'].unique() to check unique values for columns.
3. sns.countplot(y='column\_name', data=df); to visualize value counts.
4. DataFrame.duplicated().sum() to check if there any duplicated rows.

Below is one of the visualisations showing majority of students the get certified;



MISSING DATA

There is evidence of existence of missing values and they printed printed out using print(df.isnull().sum()) to see number in each column and plotted using missingno.matrix(DataFrame) as shown below;

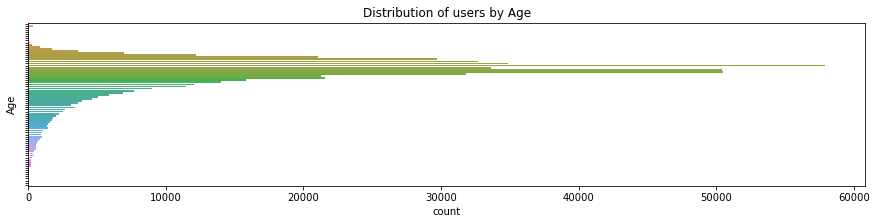
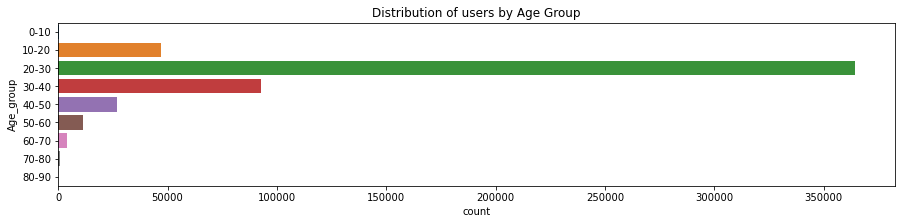


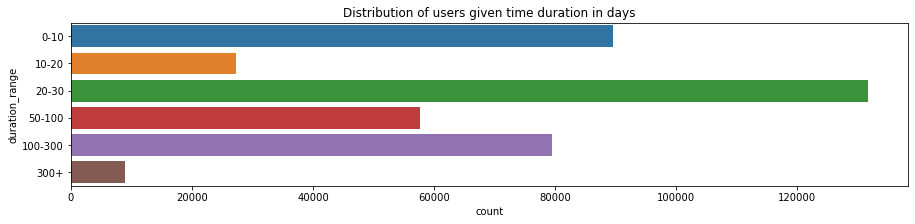
The missing values for string data types were thereafter replaced with “other” and numeric data types with the median of the variable grouped by ***course\_id*** and ***certified***.

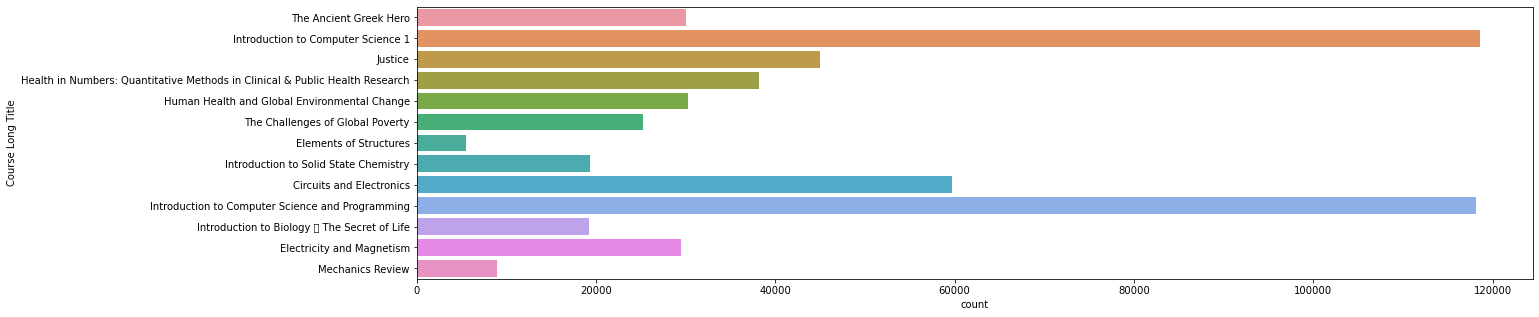
DATA STORIES AND VISUALIZATIONS

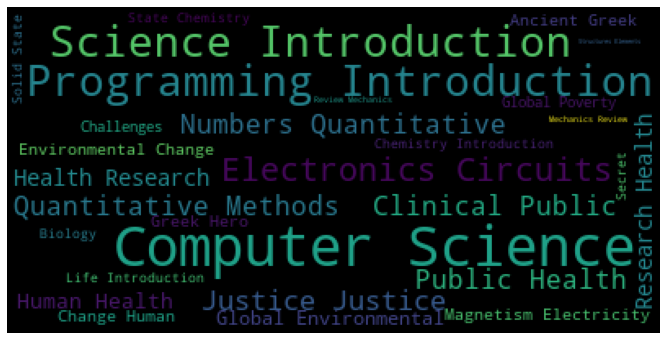
Firstly, drop the column ***last\_event\_DI*** given the new columns ***duration*** and ***duration\_range*** will be used.

Firstly, I looked at counts hence distributions of registered for different variables as seen below;

****

****

****

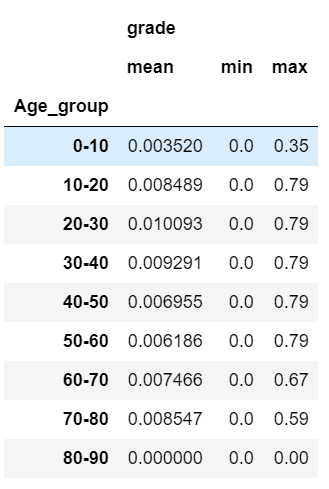
****

From the above graphs we see that registered users mainly consist of;

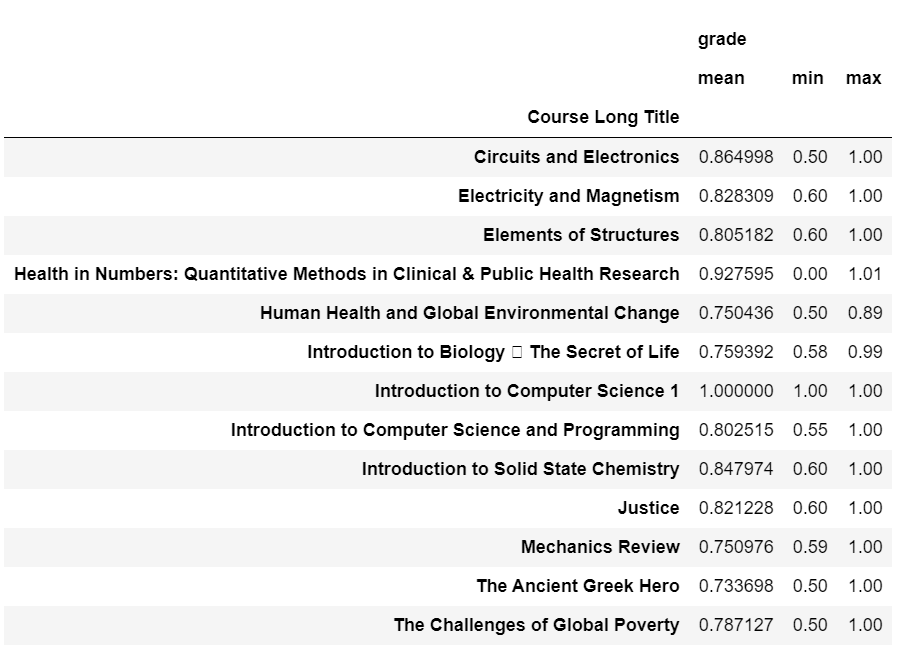
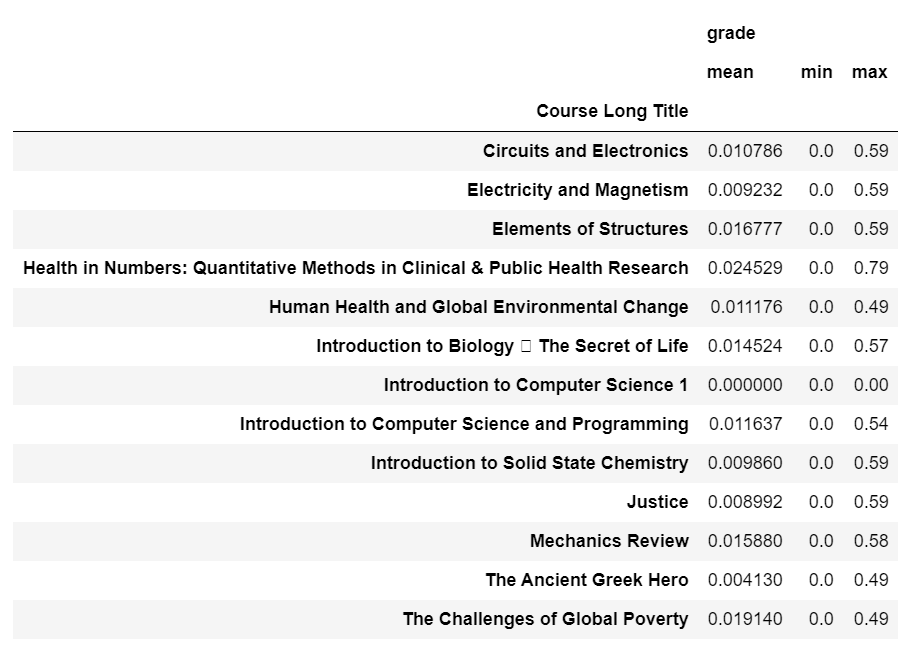
1. Age group between 20 and 30 years
2. Duration of course between 20 to 30 days
3. Both Introduction to computer science courses.
4. Males.
5. And mostly had level of education as a bachelor’s.

I then moved on to compare grades for those users that got certified and those that did not;

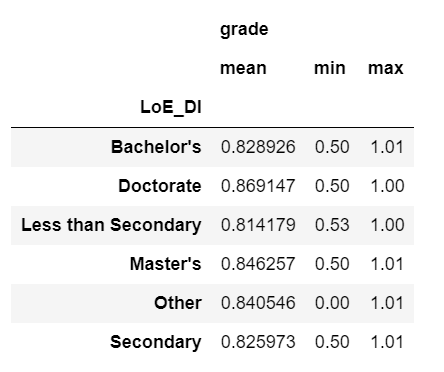
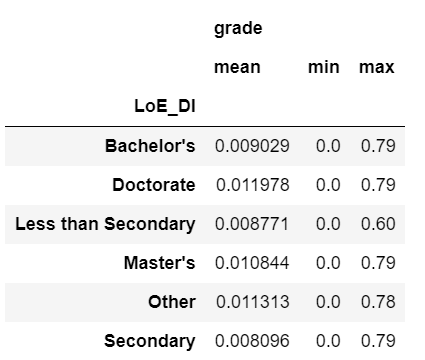
**Certified** **Not certified**

**Certified** **Not certified**

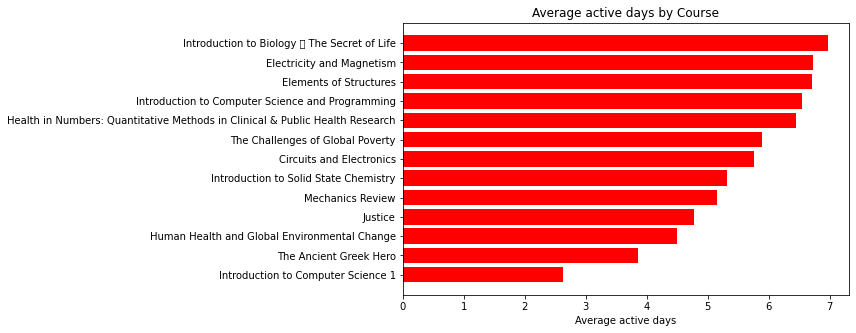
**Certified** **Not certified**

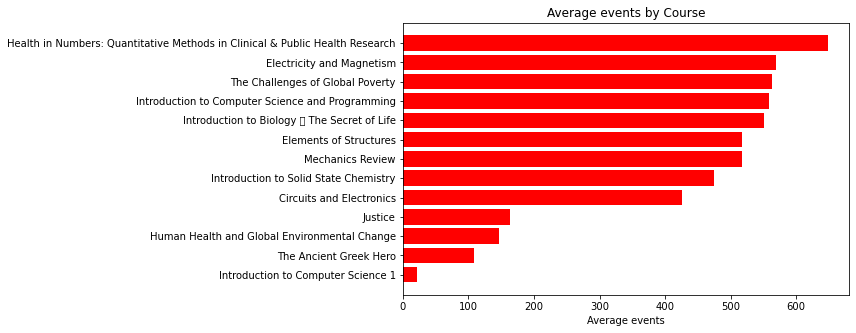
 

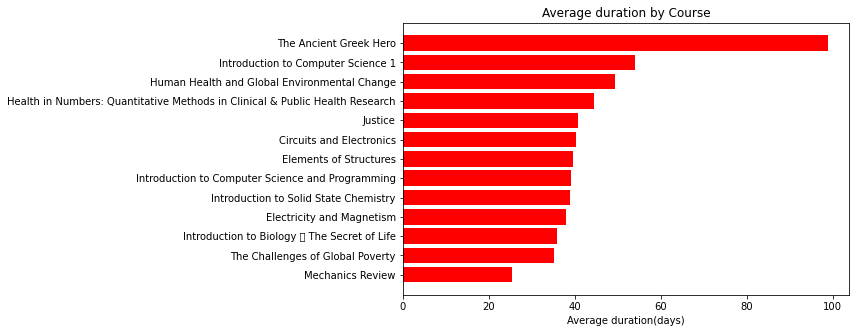
Findings from the above tables;

1. 30 to 50 year olds get the highest marks on average
2. Users doing the Health in numbers course got the highest marks on average
3. Education level does matter. Doctorate holders get the highest marks followed by Masters holders.

Thereafter we look at course activities that may have an impact on getting certified for different courses followed by the actual number of users who got certified for these courses;

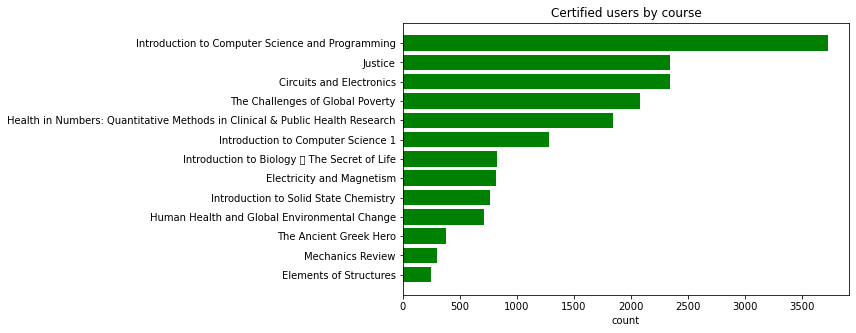


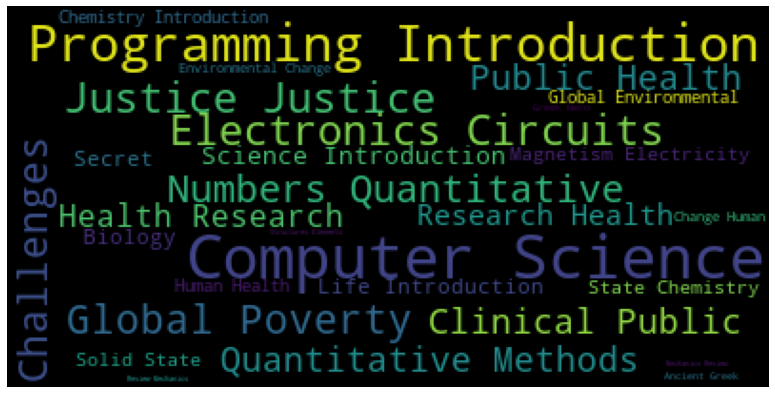
****

****

It is interesting to find the 2 courses with the least active days counts and events had the longest average duration in days. The ancient Greek hero and Introduction to computer science 1.

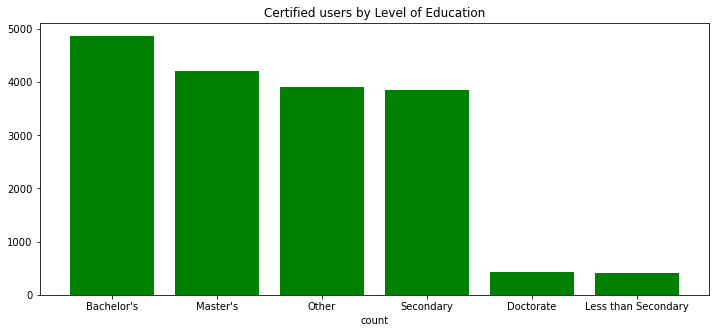
The 2 following figures look into certified users.





We see that Introduction to computer science and programming course had the most certified users.

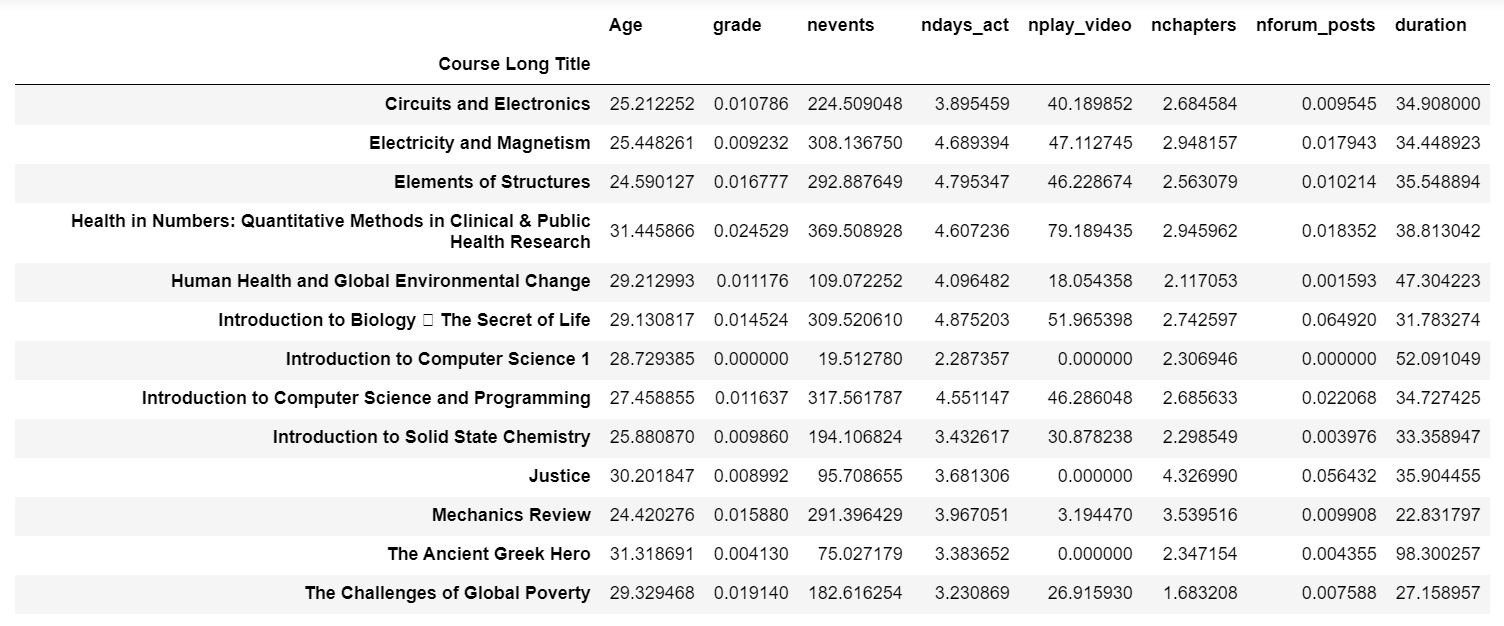
Most certified users are Bachelor’s holders;

****

The tables below show different attributes on average for certified and non-certified users per course;

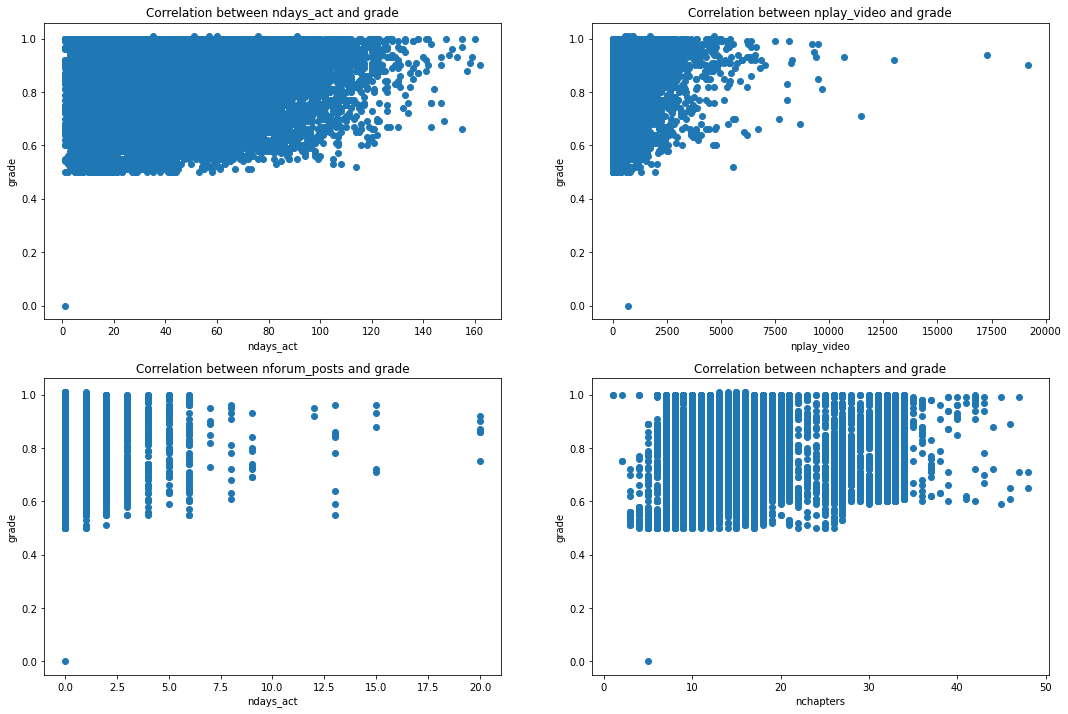


Non-certified

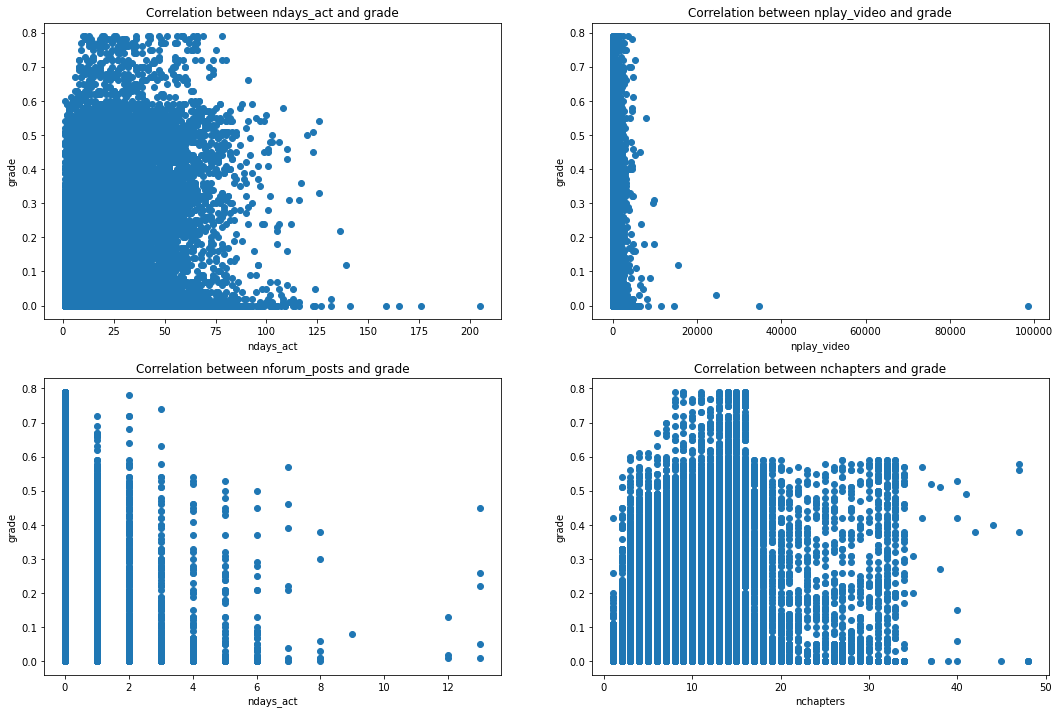


From these tables we see a clear cut that users who got certified did far more events, active days, chapters, forum posts and video plays.

Certified correlations

****

Non –certified users correlations;



No clear correlations between grades and activities.

**THIS REPORT WAS WRITTEN BY : Dalitso Chomey**

