

DATA VISUALIZATION PROJECT

PHASE ONE

1. Basic info:

Project title is: Online Courses from Harvard and MIT.

Team Work:

Name	ID	E-mail
Omar Hamed Mohamed Hamed	20191479492	CDS.OmarHamed79492@alexu.edu.eg
Mahmoud Alaa Mahmoud Abd-Elfattah	20191615074	cds.mahmoudalaa15074@alexu.edu.eg
Mohamed Moussa Mabrouk Moussa	20191613892	cds.Mohamedmoussa13892@alexu.edu.eg
Hassan Hosny Hassan Ali Khalifa	20191613617	Cds.HassanHosny13617@alexu.edu.eg
Khairy Mohamed Khairy Abdelhamed	20191480897	Cds.khairymohamed80897@alexu.edu.eg

Dataset link:

➔ [Online Courses from Harvard and MIT | Kaggle](#)

2. Overview:

- **Context:**

In 2012, the Massachusetts Institute of Technology (MIT) and Harvard University launched open online courses on edX, a non-profit learning platform co-founded by the two institutions. Four years later, what have we learned about these online “classrooms” and the global community of learners who take them?

- **Content:**

This report provides data on 290 Harvard and MIT online courses, 250 thousand certifications, 4.5 million participants, and 28 million participant hours on the edX platform since 2012.

- **Acknowledgements:**

Isaac Chuang, a professor at MIT, and Andrew Ho, a professor at Harvard University, published this data as an appendix to their paper "HarvardX and MITx: Four Years of Open Online Courses".

3. Description of the data:

The data has collected from 'HarvardX' and 'MITx' courses in 'Edx' website which provide courses for students, and it has 23 column described as follow :

Column's name	Description
1. Institution	Names of online course holders.
2. Course Number	the unique id of each course.
3. Launch Date	the launch date of each course(year/month/day).
4. Course Title	the title of each course provided by institution.
5. Instructors	the instructors of each course.
6. Course Subject	the subject of each course.
7. Year	the last time of each course
8. Honor Code Certificates	with (1), without (0).
9. Participants (Course Content Accessed)	the number of participants who have accessed the course.
10. Audited (> 50% Course Content Accessed)	the number of participants who have audited more than 50% of the course.
11. Certified	the number of participants who have been certified from the course.
12. % Audited	the percent of the audited %
13. % Certified	the percent of the certified %
14. % Certified of > 50% Course Content Accessed	the percent of the certified with accessing the course more than 50%
15. % Played Video	the percent of playing video of course.
16. % Posted in Forum	the percent of posting in forum.
17. % Grade Higher Than Zero	the percent of grade higher than zero.
18. Total Course Hours (Thousands)	total course hours (per 1000).
19. Median Hours for Certification	median hours for certification from each course.

20. Median Age	median age of the participants in course.
21. % Male	the percent of the males participated in course.
22. % Female	the percent of the females participated in course.
23. % Bachelor's Degree or Higher	the percent of bachelor's degree of higher%

4. What will we do on data:

- **Data Cleaning:**

We'll edit columns types to the right types using pandas library in python,

We also 'll check null and duplicated values before we start to visualize it.

- **Data Visualization:**

We'll use plotly library in python to visualize our data as interactive plots, we'll use Histogram, Pie chart, Scatter Plot, Line Chart, Bar Chart & 3D plot to visualize our data interactively, we 'll also use slide buttons and sub plots features to make our plots more interactive, we'll use features of colors as best we can to encourage our visualization, we'll also put question before every plot and answer that question based on the visualization we do, we'll make our visualization as slides to make it easy to discuss.