edXlytics: Learner Engagement Analytics for edX Platform

Syed Ali
OMSCS, Georgia Institute of Technology
smali@gatech.edu

ABSTRACT
UPDATED—30 July 2017. Online education is an evolving new field. Not all practices from traditional education are going to translate directly to online education. That is, it may be difficult to certain aspects of traditional education in the online area. Early feedback from existing online education systems highlights such issues as lower completion rates, which have been acknowledged by some of the thought leaders of online education. Since online education is still maturing and transforming, online educators need help identifying issues in order to formulate resolutions. Data analytics can provide this feedback, and help online educators make better decisions to improve the online education process.

This paper explores the area of Online Education Analytics which addresses the problem of low online course completion rates. The objective of this project is to analyze the data collected for CS1301x: Introduction to Computing using Python [5] on edX platforms, and provide actionable insights to Course Instructors for enhancing the overall learning experience for students. As Angelino, Williams, & Natvig [8], propose, engaging learners early and often is the best strategy to improving overall completion rates. The goal is to identify similar patterns associated with learner engagement, as well as for decline in engagement, and modify online education accordingly.

Author Keywords
Education Analytics, Learning Analytics, Engagement Analytics, edx Insights.

INTRODUCTION
Data has been around for centuries, but with recent advancements in technology, data has become much easier to collect, store and process. According to Norris, Baer & Offerman[1], , Analytics refers to "processes of data assessment and analysis that enables us to measure, improve, and compare the performance of individuals, programs, departments, institutions or enterprises, groups of organizations, and/or entire industries". Ravishanker[2] defines analytics as "Data-driven decision making, used to inform decisions at all levels of the enterprise".

Historically, the education industry has also been accumulating a lot of data across multiple aspects: Students, Teaching & Support Staff, Institutions and Courses. Here are Analytics specific to the education field as defined by Barneveld, Arnold & Campbell[3]:

- Academic Analytics: A process for providing higher education institutions with the data necessary to support operational and financial decision-making.
- Learning Analytics: The use of analytic techniques to help target instructional, curricular, and support resources to support the achievement of specific learning goals.
- Predictive Analytics: An area of statistical analysis that deals with extracting information using various technologies to uncover relationships and patterns within large volumes of data that can be used to predict behavior and events.

OVERVIEW
With the advent of Online Education, the Industry has been excited about delivering quality education to a wide range of audiences at a lower cost. Ally[4] sighted some of the major benefits of online education for both learners and instructors. With the rollout of online education, we have seen a few disappointing results, most of which failed to meet earlier expectations. One of the disappointing metrics has been low completion rates. Willging & Johnson[6] cite the following non-personal reasons for low course completion rates:

- Course quality
- Course difficulty/ease
- Lack of interaction with Instructors & Students
- Lack of technical support

High enrollment rates highlight the learners' interest in using online education as a viable medium. The high volume of data, generated by usage of online education, could be used to identify the root cause behind low completion rates after initial high enrollment. Different types of Data Analytics, sighted in Assignment 3, could be used to observe, explain, predict and influence the completion rates.

Opensource edX provides a platform where affiliated communities can contribute to building the education data analytics. Recently, edX platform started investing in the data analytics module edX Insights. Currently, edX Insights functionality is focused on Learner Data, to be utilized by user personas like Instructor, Researcher & Analyst. To summarize, some of the current edX Insights capabilities include (edX Insights documentation [7]):

- Course Enrollment: Activity, Demographics, Geography
- Student Engagement: Course Video & Content

- Student Activity: Video, Discussion & Problem
- Student Performance: Problem, Assignment & Grade

EDX: ENGAGEMENT DATA, REPORT & ANALYTICS

Currently edX provides Data and Analytics under 2 separate sections:

- Instructor Dashboard: Some of the Reports available under the Data Download Section:
  - Learner Profile
  - Problem Response
  - Grade
  - Problem Grade
- New Insights Module: Four major areas of Insights are:
  - Enrollment
  - Engagement
  - Performance
  - Learners

The edX Instructor Dashboard doesn't provide reports specifically for the learner engagement, but reports like Problem Response & Problem Grade could be utilized to analyze the engagement. Converting the response & grade reports into engagement reports would require data processing and cleanup, which might be difficult for course instructors and designers due to lack of data analytics and visualization skills.

edX Insights provides following functionality for learner engagement

- Learner Content Engagement over time
  - Active Learners: Learners who visited at least one course content page
  - Watched a Video
  - Tried a Problem
  - Participated in Discussions
- Video Engagement by course content by Complete and Incomplete with drill down capabilities into
  - Sections
  - Subsection
  - Chapter
  - Video: With details for each 5 seconds of video, broken down by unique and replay views.

EDX: ENGAGEMENT ANALYTICS ENHANCEMENTS

Abel and Kellen[9], working with the Caliper Analytics Framework and University of Kentucky, are of the opinion that Engagement Analytics could be used to analyze the learner participation and keep the instructors informed about each learners' progress. Here are some of the Engagement Analytics that could be added to edX Insight module that will provide Course Instructor to improve Learner Engagement, which will help improve the completion rates:

- Video Engagement for Course Content Feedback
- Assignment Engagement for Course Content Feedback
- Learner Course Content Activity (web page, video, assignment & discussion) for Learner Engagement and Completion Probability
- Learner Profile Analysis for Predictive Completion Probability

edX Content Insights highlights that learners engage with different types of course content, listed in decreasing order of engagement

- Content Pages
- Videos
- Assignments: Exercise, Assignment and Test
- Discussions

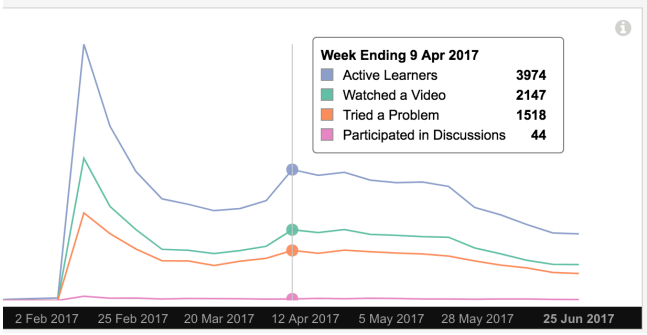

Figure 1. Learner engagement with different types of Course Content.

EDXLYTICS: LEARNER ENGAGEMENT

The edXlytics Project focusses on Learner Engagement with Video & Assignment Content. The edX Insight module is a great starting point for Video Content Engagement, but edX does not provide much functionality for Assignment Content Engagement.

Video Content Analysis

For Video Content, edX Insights modules provide good functionality for insights into learner engagement patterns. Course Instructors can interact with Video Content Analytics to understand how learners are interfacing with their course. The existing UI does have some usability limitations, however, where the Instructors need to drill down into each Section, Subsection, Chapter and Video to view the overall learner engagement.

Method

For Video Content Engagement, the edX Insight module provides descriptive analytics. The current edX Insight UI/UX requires users to aggregate some of the analytics outside the tool and stitch together observations and conclusions on their own. Using the edX Insights functionality, users get actionable insights, thus, dissolving the need to invest as much in this particular area of engagement analytics.

Observations

The overall trend in learner engagement shows decrease in engagement over the course of the class. The following Insight graph shows learner engagement specific to video content. All of the engagement numbers indicate that the video consumption reduces with every unit. The largest number of 'incompletes' (least engagement) is for the

Course Information Unit, and the sharpest engagement decline is after Unit 1.

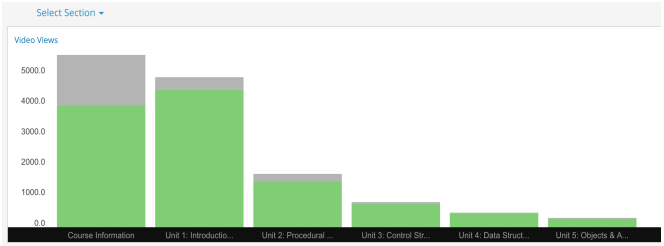

Figure 2. Learner engagement with Video Content. Engagement reduces with every unit. Course Information has highest incomplete rate. Sharpest decline is after Unit 1.

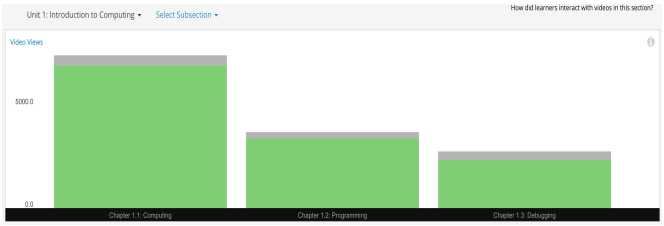

Figure 3. Learner engagement with Video Content for Unit 1: Introduction to Computing. Engagement reduce with every subsection.

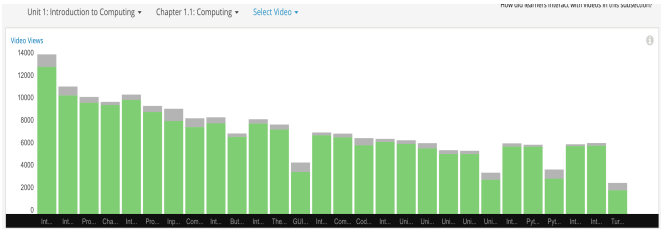

Figure 4. Learner engagement with Video Content for Unit 1: Introduction to Computing and Chapter 1.1: Computing. Engagement reduce with every video.

Video Content Engagement Insights provide Instructors a drill down functionality into each Section, Subsection and Chapter. This drill down functionality is especially useful to identify irregularities in the pattern of engagement. The graph below highlights one of these irregularities in Unit 1: Introduction to Computing, Chapter 1.2: Programming and Chapter Preview (1.2.1.2), which has a lower engagement or lower completion rate compared to other views in the same chapter. This information is actionable for the course instructors and designers to analyze those videos and content specifically highlighted, and to update them for improved views and completion rates.

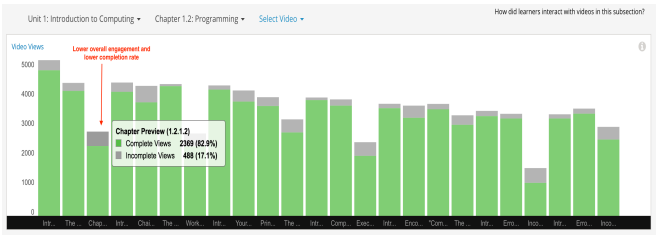

Figure 5. Learner engagement with Video Content for Unit 1: Introduction to Computing and Chapter 1.2: Programming. Chapter Preview (1.2.1.2) has significantly lower views and lower completion rate.

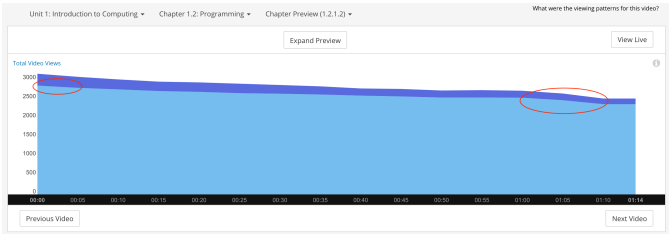

Figure 6. Learner engagement with Video Content for Unit 1: Introduction to Computing, Chapter 1.2: Programming and Chapter Preview (1.2.1.2) video. Engagement reduces with time, but rate of decline is sharper between seconds 0-5 and 1:00-1:10.

Conclusion

Video Content Insights provide descriptive analytics for the video content. These descriptive analytics can be used by course designers and instructors to take action and improve leaner engagement, even prevent learners from dropping out. There is some scope for Video Content Engagement Insight, but for this particular project it is a lower priority to focus on.

Assignment Engagement Analytics

Since neither the edX Insights nor the Instructor Dashboard provide functionality for Assignment Engagement Analytics, the edXlytics Project investigated Assignment Engagement data in order to extract actionable insights for Course Designers and Instructors.

Method

For Assignment Engagement, the Problem Grade Report from the edX Instructor Dashboard | Data Download Section was the starting point. The Problem Grade Report contains performance data for each learner for each Assignment. Assignments could be one of the 3 types:

- Exercise
- Problem Set
- Test

Using Python scripts, Assignment Performance data was converted to Assignment Engagement data. Different types of Assignment data was processed and analyzed separately. Once the data was cleaned and formatted, it was imported it

into MySQL databases for filtering, sorting and aggregating. Analyzed data was visualized using excel charts and python Seaborn library plots.

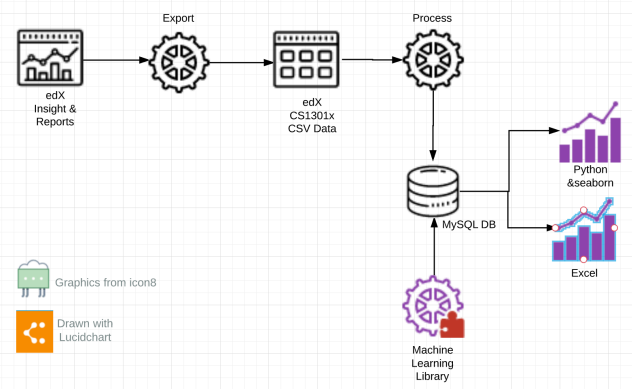

Figure 7. Design for the Data Pipeline for Assignment Engagement Analytics.

Observation: Declining Assignment Engagement

Assignment Engagement declines over the course of the class for all Assignment types: Exercise, Problem Set and Test.

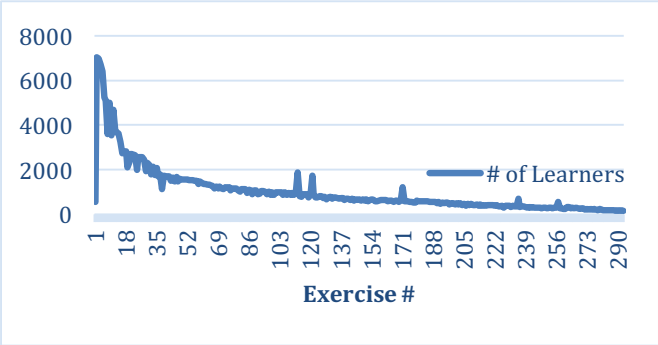

Figure 8. Learner engagement with Exercise Assignments. Total 294 Exercises in CS1301x Course.

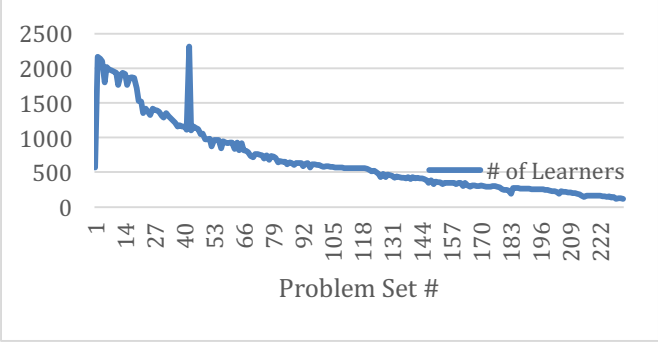

Figure 9. Learner engagement with Problem Set Assignment. Total 233 Problem Set in CS1301x Course.

Since Assignment Engagement trends are continuously declining, it is difficult to identify the Assignments which might correlate to higher drop off rates. Different types of analyses were researched to identify correlation between Assignment and engagement.

The following four analytics were used:

- Rate of Change
- Rate of Rate of Change
- Delta from Moving Average
- Bollinger Band

Observation: Identifying Assignments with higher drop-off

Since engagement graphs had fluctuations, the Rate of Change highlighted many assignments that could be related to the dropout rate. The Rate of Rate of Change smoothed the graph too much and didn't provide enough insights. Delta from Moving Average was the most intuitive in highlighting the Assignments that were related to the highest drop offs.

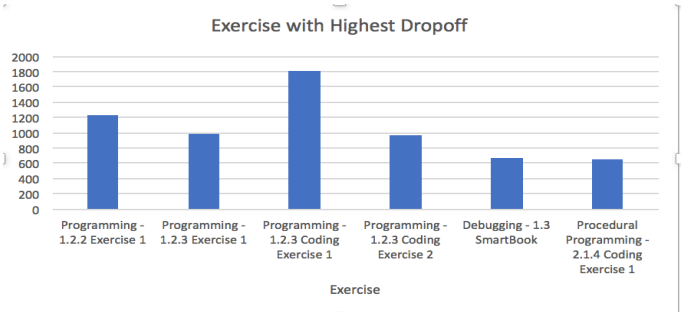

Figure 10. Exercise with Highest Drop-off

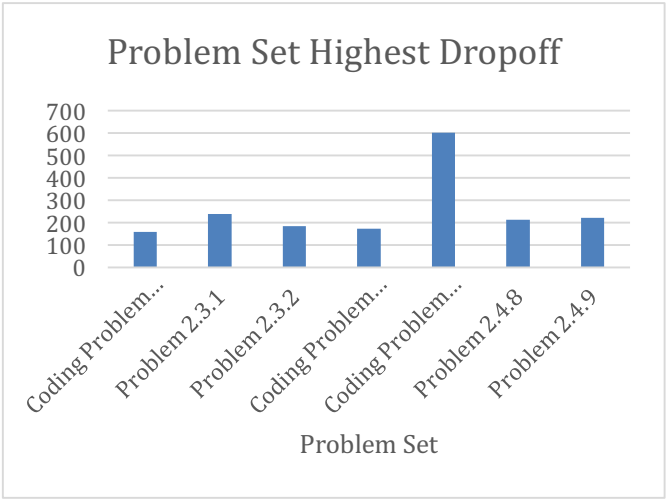

Figure 11. Problem Set with Highest Drop-off

Observation: Identifying Assignments with higher completion

Since engagement graphs had fluctuations, using moving averages and Bollinger Band Graphs one can highlight the parts of the graph where the changes were more drastic. For

a continually decreasing graph, such as that of user engagement, Bollinger Band Graphs are able to highlight the change in trend. For the Exercise Engagement graph, Bollinger Band highlights surge in engagement. Growth in engagement was seen with the following Exercises.

- Loops - 3.3.2 Exercise 1
- Loops - 3.3.4 Exercise 2
- Error Handling - 3.5.4 Exercise 2
- File Input and Output - 4.4.3 Exercise 2
- Dictionaries - 4.5.4 Exercise 2

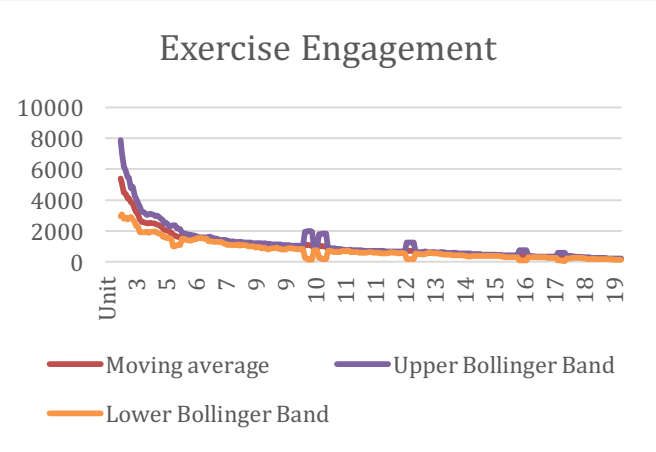

Figure 12. Exercise Engagement with moving averages and Bollinger Band

This analysis highlights the exercises that draw higher engagement, compared to the other exercises in the section and subsections. However, it doesn't provide an insight into the cause of the increased engagement. Course Designers and Instructors will need to analyze the Assignment content to explain the uptick in engagement. Once these relationships are better understood, Designers & Instructors can incorporate the date into other Assignments to improve engagement.

Observation: Coding Problem Set has worst completion rate

Engagement for the Problem Set is higher than engagement for the Coding Problem Set. Specifically, engagement is 20-30% higher for the non-coding problem set. Here again, designers and instructors need to figure out why learners are not engaging with coding problem sets even though this course is all about learning how to code with Python. The spike in engagement for Unit 3 merely accounts for Multiple Choice questions. We also see higher engagement rates for the Multiple Choice Question Problem Set.

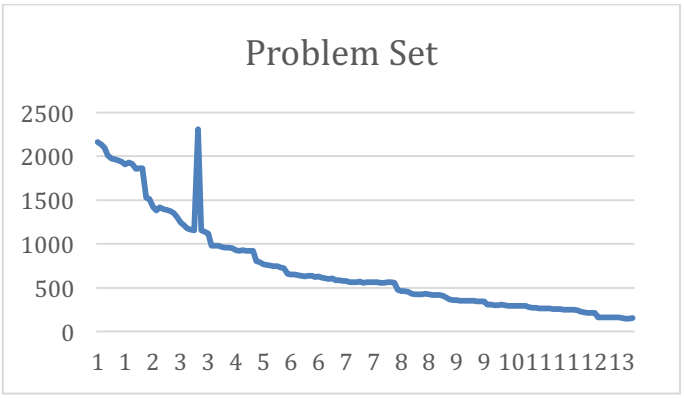

Figure 13. Problem Set Engagement, excluding Coding Problem Set. Spike in Unit 3 is for a multiple choice question.

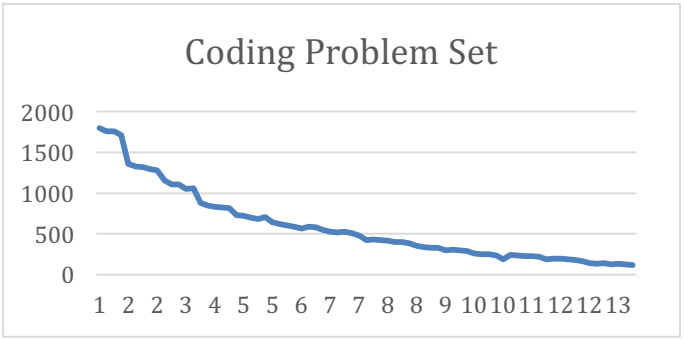

Figure 14. Coding Problem Set only Engagement. Engagement is 20-30% lower than non-coding Problem Set.

Observation: Higher drop off after major milestone

Analyzing the Assignment Engagement data reflects that learner engagement continues to drop as the course progresses, but the rate of drop is higher around the milestones in the course. That is, drops are steeper when the course transitions from one section, subject, or chapter to another. We see the trend consistently for Exercise and Problem Set Assignment Types. This observation is also consistent with the Video Content Engagement from edX Insights.

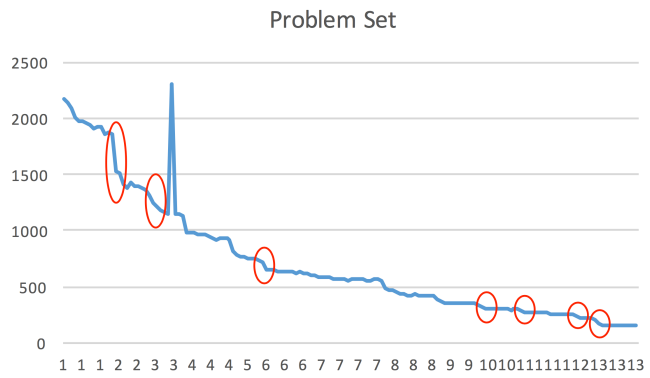

Figure 15. Problem Set drop-off around course milestones.

Analyzing the content for the final sections in Unit 1, (Chapter 1.3 "SmartBook" & Chapter 1.3 "Additional Resources"), it is seen that the content gives the learner accolades for completing the first unit. This kind of milestone seems consistent with traditional education where instructors reward learners for persistence. With MOOCs, too, it is important to motivate learners. Perhaps even more important, in MOOCs situations, is the need to keep learners engaged with future assignments by creating a hook into the next unit. As with television shows, when an episode concludes previews of the next episode are shown to pique curiosity and increase engagement. In the online education arena, instructors could give a simple snapshot of what follows. A hook such as this could benefit online courses by keeping leaners engaged.

Learner Engagement Score & Profile Analysis

Next, the edXlytic project continued investigating learner engagement specific to assignment content. The method of this analysis involved finding correlation between learners' engagement scores and different aspects of learner profiles.

Method

Analyzing learner engagement and various profiles started with Learner Profile and Problem Grade data from the Instructor Dashboard, Data Download section. Using a python script, Problem Grade Data was converted to Learner Assignment Engagement data. Then, the assignment engagement score was calculated for each learner using the following two methods:

1. Counting the number of assignments attempted, and calculating the percentage out of the total number of assignments in the course.
2. Calculating the furthest progression through entire syllabus of assignments in the course, and calculating the percentage out of total number of assignment in the course. (This method was used for further analysis).

The engagement scores were calculated separately for two different types of assignments: Exercise and Problem Set. Then, the total score for each learner was calculated by taking the average engagement score for Exercise and Problem Set engagement. Next, the newly calculated learner assignment engagement score was joined with the learner profiles to see trends by different attributes of specific profiles. Here are the learner profile attributes that were used to analyze the engagement score:

- Gender
- Year of birth
- Level of education
- Country

After joining the learner engagement score with the learner profile attributes, data was run through different filtering and sorting mechanisms, and plugged into data visualization charts. Initially, excel charts were used for data visualization. Eventually, Python scripts (using Seaborn, pandas & Matplotlib libraries) were used for more comprehensive analyses.

Observations: Gender

Gender doesn't seem to have any correlation on the learner population or the average learner engagement score by gender. In the following bar chart we see that the average engagement score varies by within 2% for different genders.

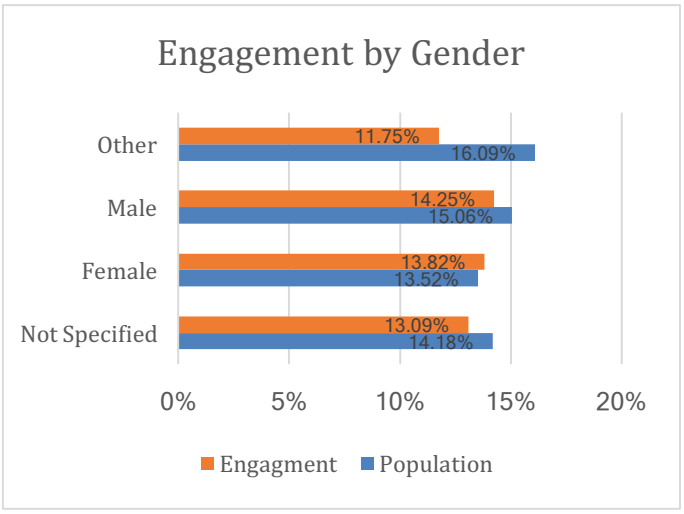

Figure 16. Average Learner Engagement Score by gender. Only 2% variation by gender.

Figure 17 This box and whiskers chart also confirms the same observation that the median and quartile values are consistent for different values of gender. Even though there are more Male learners compared to Females, Not Specified and Other, the engagement score ranges are fairly uniform.

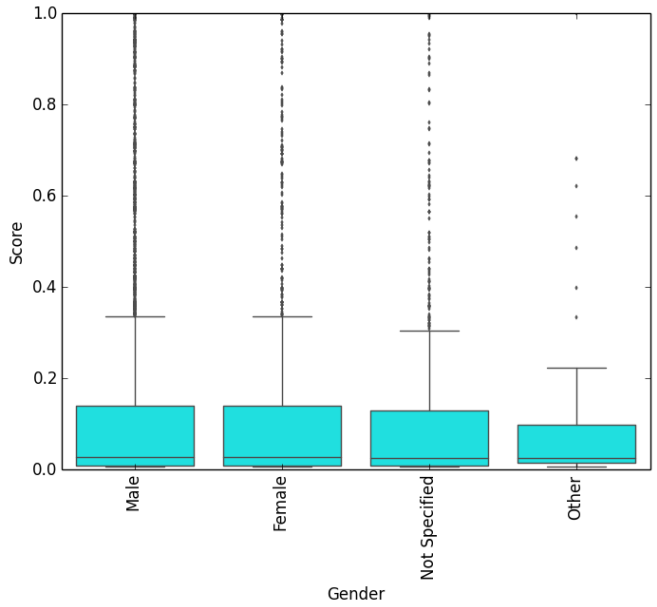

Figure 17. Box and whisker chart for Learner Engagement Score by gender.

Observations: Age

Figure 4. depicts the distribution of average learner engagement scores by the decade in which the learner was born. The engagement score is highest for learners born in the 1940s, and gradually decreases for younger learners born later. The pattern is not consistent for learners born in the 1900s, 2010s and 2020s, which are also outliers and can be considered as invalid data. After all, it is hard to imagine Python learners that are

- older than 110 years old (1900s)
- younger than 7 years old (2010s)
- not born yet (2020s)

The Learner Count line in Figure 18 clearly shows that the number of learners born in the 1900s, 2010s & 2020s are really small, and could be considered as outliers. Outlier data was filtered out for further analysis of Learner Engagement Score and Age.

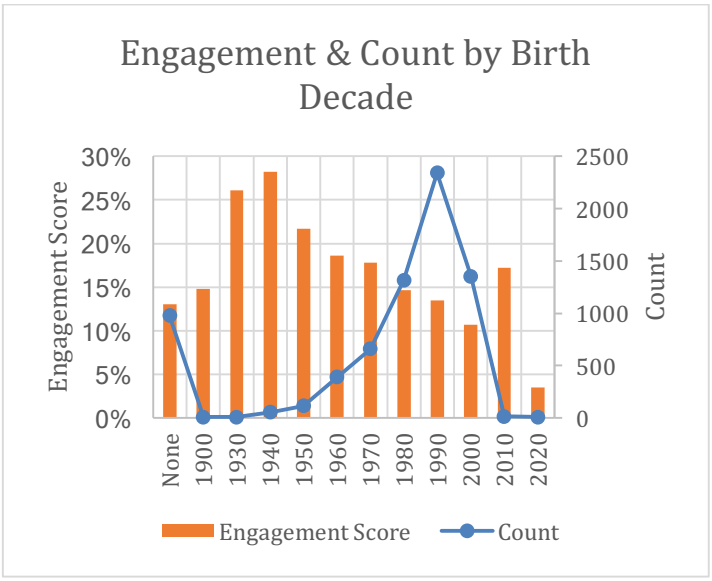

Figure 18. Bar Chart for Learner Engagement Score by Birth Decade, super imposed with Learn Count by Birth Decade.

To further analyze the relationship between learner birth year and learner engagement, a linear regression analysis was performed. A scatter plot was made plotting learner birth year against learner engagement score. A regression analysis was also applied to plot the trend line. In Figure 19 the trend line clearly shows a negative correlation between learner engagement and learner age.

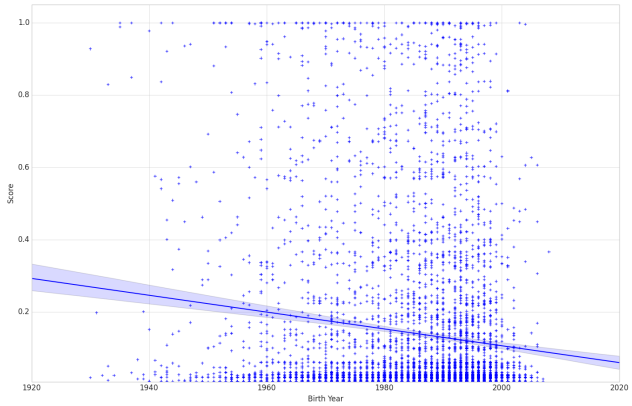

Figure 19. Scatter Plot for Learner Engagement Score and Birth Decade, with regression analysis trend line.

A Correlation Analysis between Learner Birth Year and Learner Engagement Score was performed, the results of which are shown in Table 1.

	Birth Year	*Engagement Score*
Birth Year	1	
Engagement Score	-0.120259825	1

Table 1. Correlation Analysis for Learn Birth Year and Learner Engagement Score.

Observations: Level of Education

The level of education seems to have a correlation to the learner engagement rates. Learners with Doctorate degrees have the highest engagement. Learners with Master's & Bachelor's degrees have the next highest engagement.

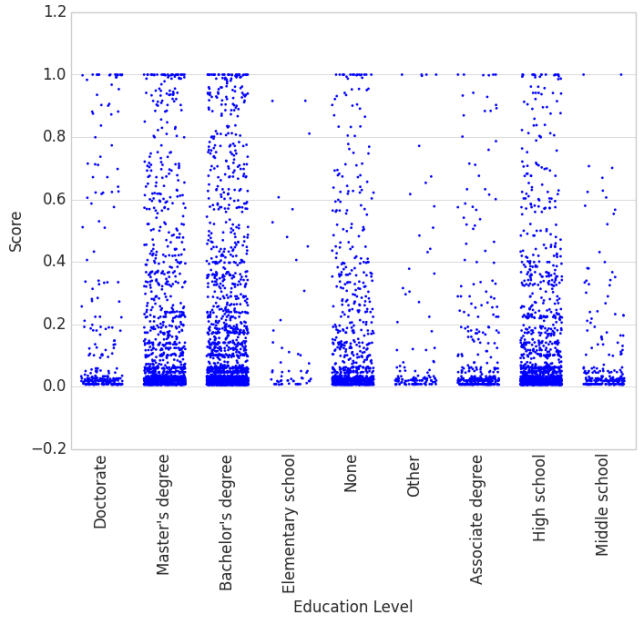

Figure 20a. Point Scatter Chart for Learner Engagement Score by Learner Education Level, sorted by highest mean Engagement Score.

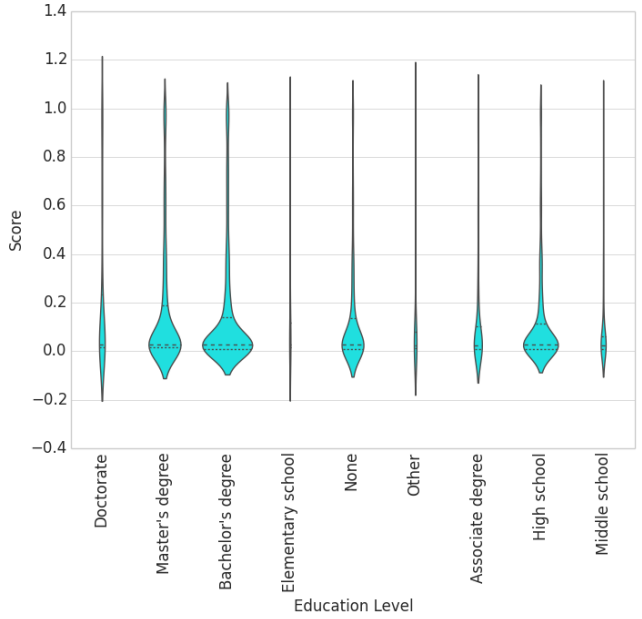

Figure 20b. Violin Chart for Learner Engagement Score by Learner Education Level, sorted by highest mean Engagement Score.

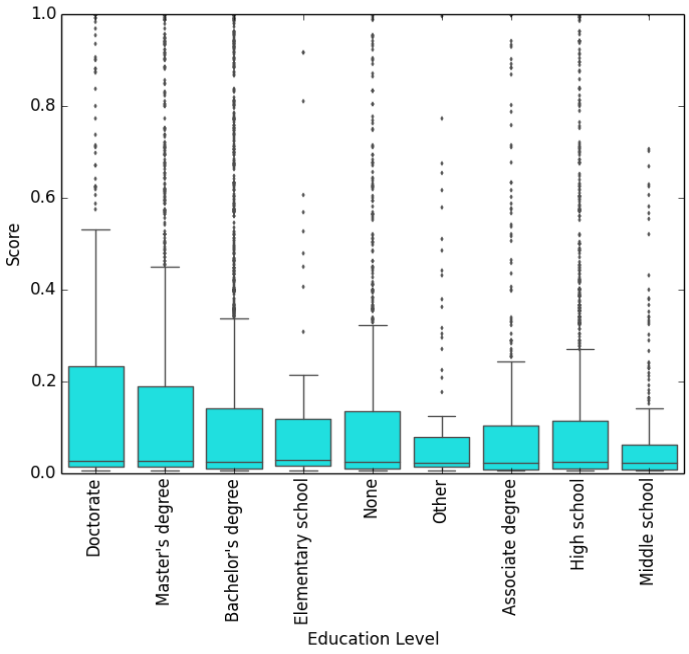

Figure 20c. Box and Whisker Plat for Learner Engagement Score by Learner Education Level.

Figures 20a, 20b & 20c show distribution of learner engagement categorized by Learner's previous education level. Figures are sorted in order of decreasing average engagement. Figure 20c shows that the highest number of engaged learners already possess Bachelor's Degrees, Master's Degrees and High School Degrees. However, the highest average engagement score is of learners with Doctorate Degrees. Another interesting observation is that the median engagement score was completely flat for different levels of education, even though mean engagement scores declined as the education levels decreased.

Observations: Country

In order to analyze the learner engagement data from 162 different countries, I analyzed the data from 3 different vantage points:

- Top 10 Average Engagement Countries
- Lowest 10 Average Engagement Countries
- Top 10 Total Enrollment Countries

Figure 8 shows the Top 10 Countries with the highest average learner engagement, along with the number of engaged learners from those specific countries. Most of the top 10 most engaged countries are in Europe, but a couple are in Asia (Taiwan & Bangladesh), and 1 in South America (Chile).

Figures 21a, 21b & 21c show different plots for the top 10 Countries with the highest average engagement. The plots show the engagement score distribution for different countries in decreasing order, however, there are no obvious patterns in median and quartile variations for those countries.

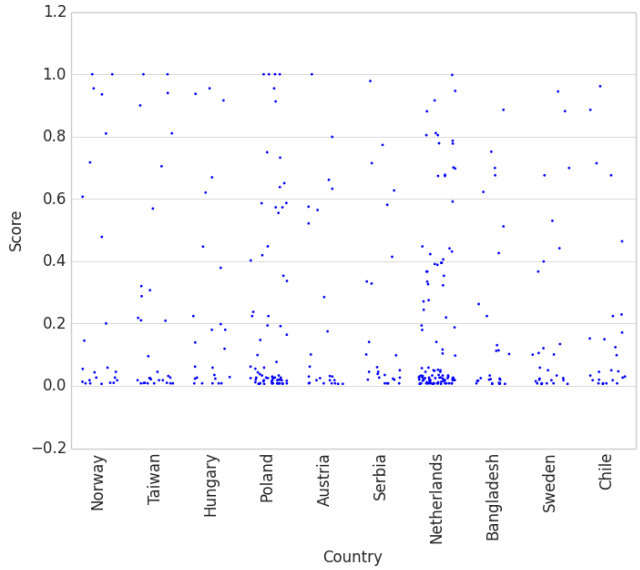

Figure 21a. Scatter Plot for Learner Engagement Score by Country, for Countries with Top 10 highest mean engagement.

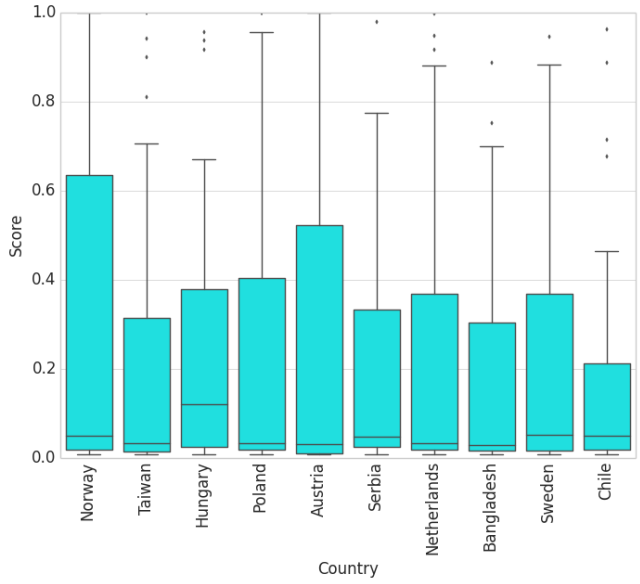

Figure 21b. Box and whisker Plot for Learner Engagement Score by Country, for Countries with Top 10 highest mean engagement.

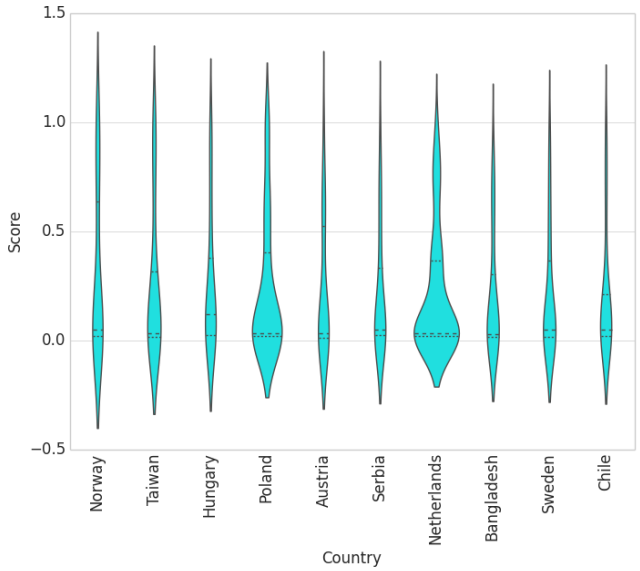

Figure 21c. Violin Plot for Learner Engagement Score by Country, for Countries with Top 10 highest mean engagement.

Figures 22a, 22b & 22c show the bottom 10 countries with the Lowest average learner engagement. The plots show the engagement score distribution for different countries in decreasing order.

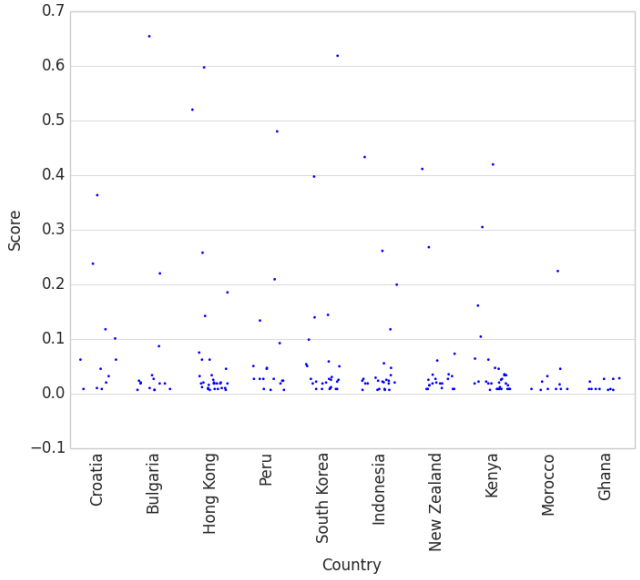

Figure 22a. Scatter Plot for Learner Engagement Score by Country, for Countries with Bottom 10 lowest mean engagement.

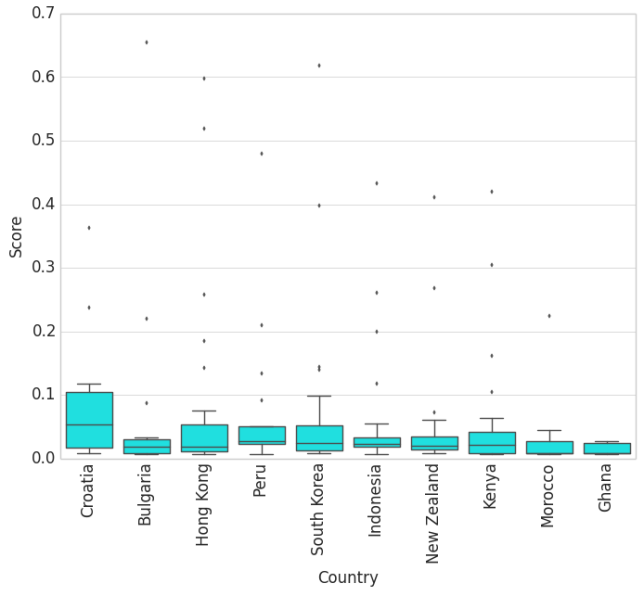

Figure 22b. Box and whisker Plot for Learner Engagement Score by Country, for Countries with Bottom 10 lowest mean engagement.

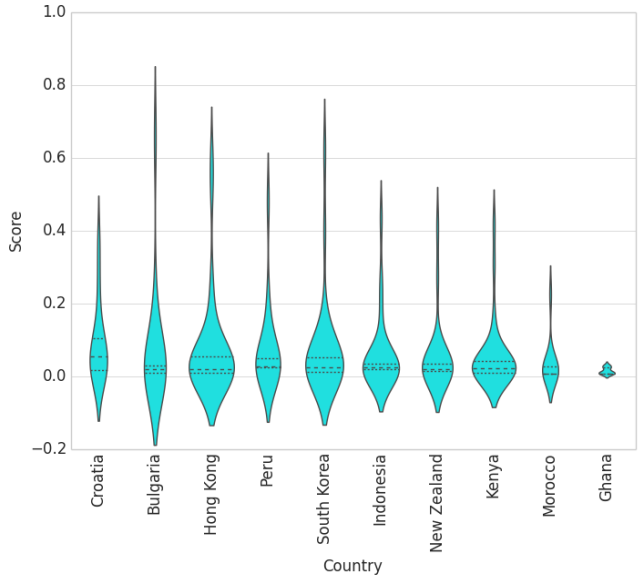

Figure 22c. Violin Plot for Learner Engagement Score by Country, for Countries with Bottom 10 lowest mean engagement.

Figures 23a, 23b & 23c show Learner Engagement for Countries with highest total enrollment. The plots show the engagement score distribution for different countries in decreasing order. Figure 23b plot shows that the median engagement score is consistent for countries with high numbers of engaged enrolled learners.

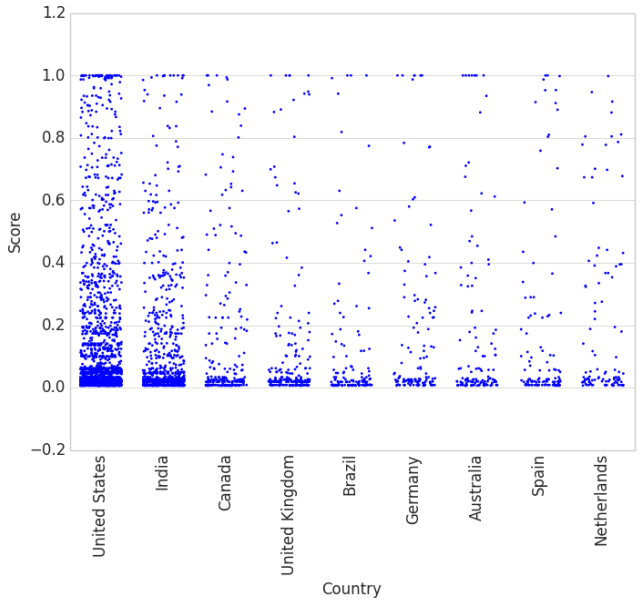

Figure 23a. Scatter Plot for Learner Engagement Score by Country, for Countries with highest enrollment.

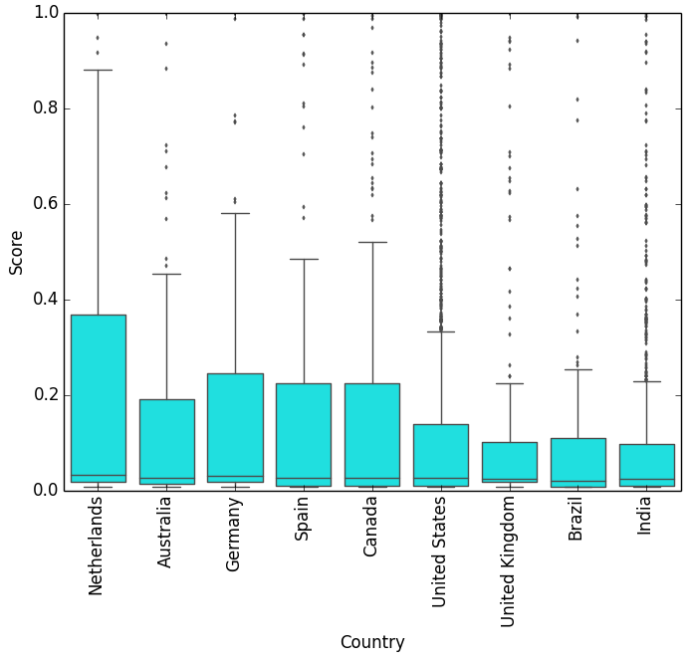

Figure 23b. Box and whisker Plot for Learner Engagement Score by Country, for Countries with highest enrollment.

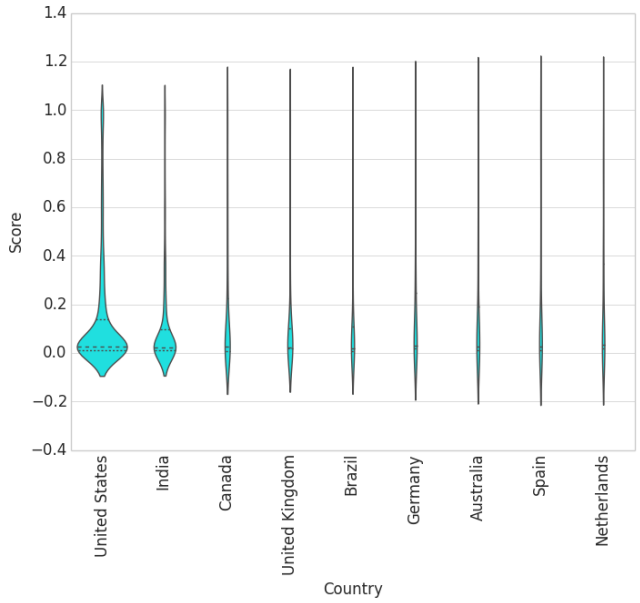

Figure 23c. Violin Plot for Learner Engagement Score by Country, for Countries with highest enrollment.

Analyzing the learner data by country, there was no evidence of any substantial correlation with the learner engagement.

Conclusion

In studying the Learner Engagement Analysis, the following inferences could be made about Learner Profile Attributes:

- Gender: Learner Engagement is NOT dependent on Learner Gender
- Year of birth: Learner Engagement is negatively correlated to a Learner's Birth year. That is, Learner Engagement goes down, for more recently born or younger Learners.
- Level of Education: Learner Engagement is positively correlated to Learner's Level of Education. That is, Learner Engagement goes up as the Level of Education increases.
- Country: There is no obvious correlation between Learner Engagement and Learner's Country.

CONCLUSION

edX Video Content Insights provides descriptive analytics which could be used by course authors and instructors to improve the content for better learner engagement. Assignment Content Analysis provided great insight into which assignments kept the learners engaged and which ones caused sharp drops in engagement. Learner Engagement Score & Profile Analyses highlighted the Profile attributes that could be related to engagement. All of these analytics can provide course authors instructors valuable insight into aspects of course and learner profiles that could be affecting the engagement.

NEXT STEPS

The next step would be to plan and organize user testing with these analytics and actionable insights. These analytics may require some enhancement for users to get the maximum value to improve learner engagement. Once the design of these analytics is understood easily, the next step would be to enhance the edX Analytics Pipeline to aggregate this engagement data into a result store, so that it could be visualized using Insight API and UI.

ACKNOWLEDGMENTS

I am really grateful to my mentor Ken Brooks, who provided valuable pointers for providing better data visualization for different analysis. My previous mentor Habib Khan was also very helpful in enabling me to get familiar with Education Technology field.

REFERENCES

1. Donald Norris, Linda Baer, and Michael Offerman (September, 2009). A National Agenda for Action Analytics. Retrieved May 27, 2017, from http://lindabaer.efoliomn.com/uploads/settinganationalagendaforactionanalytics101509.pdf
2. Ganesan Ravishanker (February 2011). Doing Academic Analytics Right: Intelligent Answers to Simple Questions. Retrieved May 27, 2017, from https://library.educause.edu/resources/2011/2/doing-academic-analytics-right-intelligent-answers-to-simple-questions
3. Angela van Barneveld, Kimberly E. Arnold, and John P. Campbell (January 2012). Analytics in Higher Education: Establishing a Common Language. Retrieved May 27, 2017, from https://library.educause.edu/~/media/files/library/2012/1/eli3026-pdf.pdf
4. Mohamed Ally (2004). FOUNDATIONS OF EDUCATIONAL THEORY FOR ONLINE LEARNING. Retrieved May 27, 2017, from http://cde.athabascau.ca/online_book/ch1.html
5. David Joyner (Feb, 2017). Introduction to Computing using Python Retrieved June 3, 2017, from https://www.edx.org/course/introduction-computing-using-python-gtx-cs1301x
6. Pedro A. Willging & Scott D. Johnson (OCTOBER 2009). FACTORS THAT INFLUENCE STUDENTS' DECISION TO DROPOUT OF ONLINE COURSES. Retrieved May 27, 2017, from http://files.eric.ed.gov/fulltext/EJ862360.pdf

7. edX Insights documentation. Overview of EdX Insights. Retrieved May 27, 2017, from http://edx.readthedocs.io/projects/edx-insights/en/latest/Overview.html
8. Lorraine M. Angelino, Frankie Keels Williams & Deborah Natvig (Jul 2007). Strategies to Engage Online Students and Reduce Attrition Rates. Retrieved June 7, 2017, from https://eric.ed.gov/?id=EJ907749
9. Rob Abel & Vince Kellen (2015). Simplifying Learning Analytics via the Caliper Analytics Framework. Retrieved July 1, 2017, from http://www.educause.edu/sites/default/files/library/presentations/E15OL/OL01/OL01_Simplifying%2BLearning%2BAnalytics.pdf
10. Vivek Murali (October 2014). Diving into Data Analytics Tools in K-12. Retrieved May 27, 2017, from https://www.edsurge.com/news/2014-10-06-diving-into-data-analytics-tools-in-k-12
11. George Siemens and Phil Long (2011). Penetrating the Fog: Analytics in Learning and Education. Retrieved June 3, 2017, from https://eric.ed.gov/?id=EJ950794