Analysis of Users by Geographical Location

Introduction

EdX is one of the first learning platforms that is truly used globally. The data on completion rates and user demographics includes information about geographical location, which is an especially interesting variable to analyze because different countries also have different cultures, educational systems, economies, and levels of opportunity.

Putting a large dataset in front of stakeholders and attempting to convince them of any findings would be a difficult and unnecessary task. Creating visualizations helps establish clear discoveries in our analysis. In our case, we plan to create geographical heat maps to show our findings across different regions.

In order to understand the cultural and economic differences that might cause different levels of engagement and completion across the globe, we turned to the UN's human development index. The human development index is a measurement of development in countries based on health and life expectancy data, education and knowledge data, and standard of living and economic data.

There is one notable study that has been done on the topic of cultural differences and MOOCs. Researchers at North Carolina State University looked at how geographic and cultural factors might influence student use of and performance in MOOCs. They highlighted the Hofstede Dimensions, six factors such as individualism and masculinity that describe each country, and made inferences on how those cultural differences might impact student use.

Learning objective

In our project, we focused on analyzing how users from different regions interacted with the online courses and attempt to answer the following question: do users across the globe interact with online courses in the same way and are certain regions more likely to have users get certified than others? To do these we analyzed the following parameters by region (country):

1. Total number of users who explored the course, and proportion with respect to the total population
2. Among users who explored the courses, what percentage of them got certified?
3. Among users who explored a certain course, what percentage of them got certified?

4. How did users between regions differ in terms of how many total events and events per day they interacted?
5. Relationship between all these factors and the Human Development Index

User and context

As EdX is one of the leading sources of online, internationally accessible education, it would be valuable for users to better understand how the product is reaching users from different countries. Also, further analysis of how international users are interacting with the platform might allow course designers to better reach all users. For example, this analysis might show that video watching rates are low for a certain region or country, and further investigation might indicate that this is due to slow internet speeds, or that schools in that country rarely show videos so users are not used to learning from videos. This study might prompt course designers to change the way their course is designed to better reach those users.

This analysis could be useful to course administrators and stakeholders in edX who desire to make learning equitable across all ranges of people. They would care about countries that have lower completion completion rates and less people accessing/interacting with the courses. They may ponder as to what challenges/barriers are certain people facing that make completing and accessing the course more difficult for them.

Design and method

We are first going to use R to gather the summary statistics we would like to showcase on our geographical heat maps. We will first clean and filter the data using a library called dplyr. After filtering the data by region and users who explored their respective courses, we will look at our desired summary statistics from this subgroup to get an overall picture. We will then further refine the data and filter by course. The three courses we chose to filter were HarvardX/CS50x/2012, MITx/8.02x/2013_Spring, and HarvardX/ER22x/2013_Spring.

With these courses, we felt that we had a broad set that were a good representative of all the courses in the data set. The three courses were a computer science course, a physics course, and a justice course. We did summary analyses with each of these courses applied to our region and explorer filtered dataset. At the end of our data analysis we were left with all of the following statistics for each region:

Total number of users enrolled, normalized by population
Human Development Index
Certified Rate from the users who explored courses overall
Certified rate from users who explored CS250
Certified rate from users who explored 8.02

Certified rate from users who explored Justice
Number of actions per user
Average Actions Total
Average Actions per Day

After finishing up our data analysis and getting our summary statistics, we used Carto, an online mapping software, to display our findings.

Results

We were able to create eight different maps of our data, shown below. There is also a link to further explore our maps. Due to a limitation of carto, all of the layers are displayed when the explorer is opened, but to look at an individual map, the checkbox next to its legend must be checked and all other checkboxes must be unchecked. We colored the maps by category in order to better help the user ensure they are looking at the correct map, with the population map in red, human development index in orange, all certification rate maps in blue, and number of actions maps in purple.

https://mit.carto.com/u/katief/builder/49d63982-f24a-43ff-a9f0-e613edf1ee13/embed

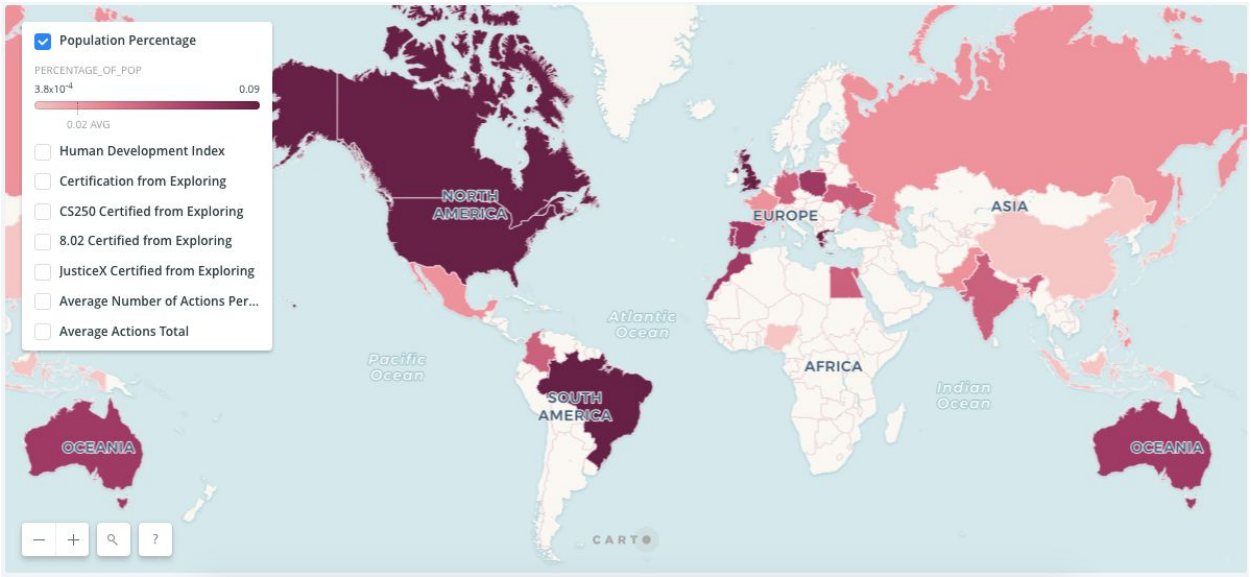

Figure 1: Percentage of Country's Population using EdX

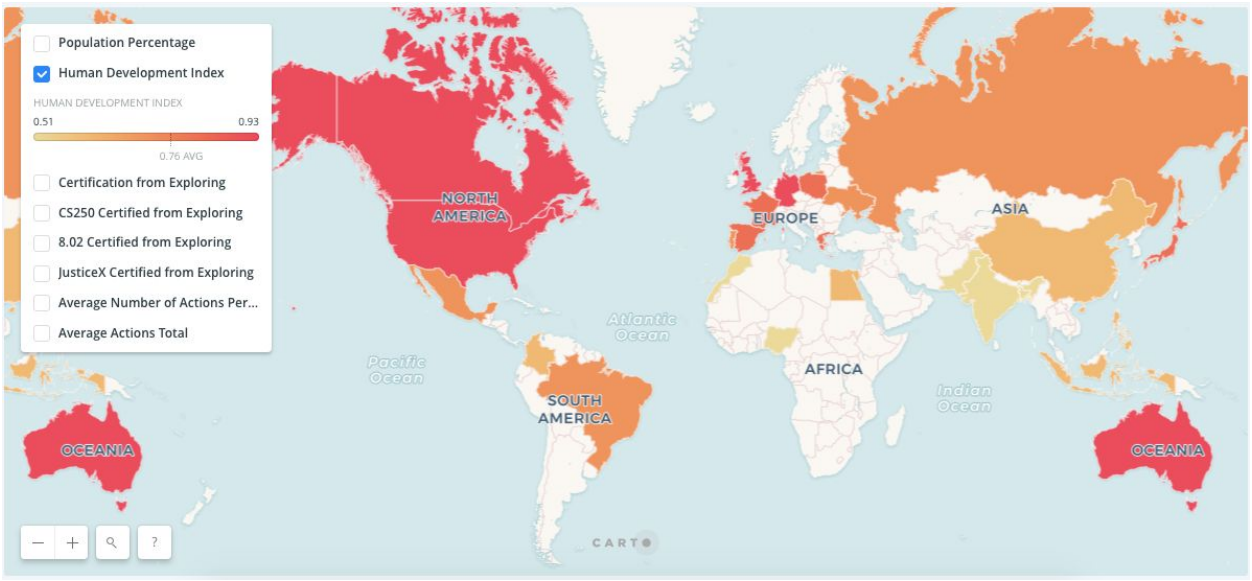

Figure 2: Human Development Index

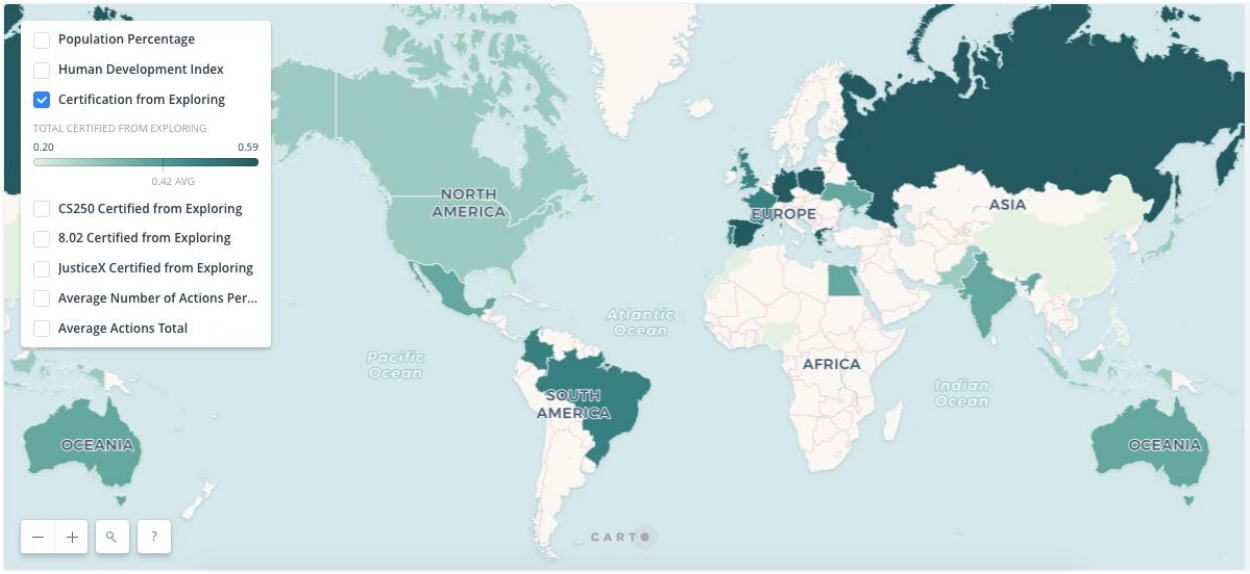

Figure 3: Rate of Certification from Exploring across all Courses

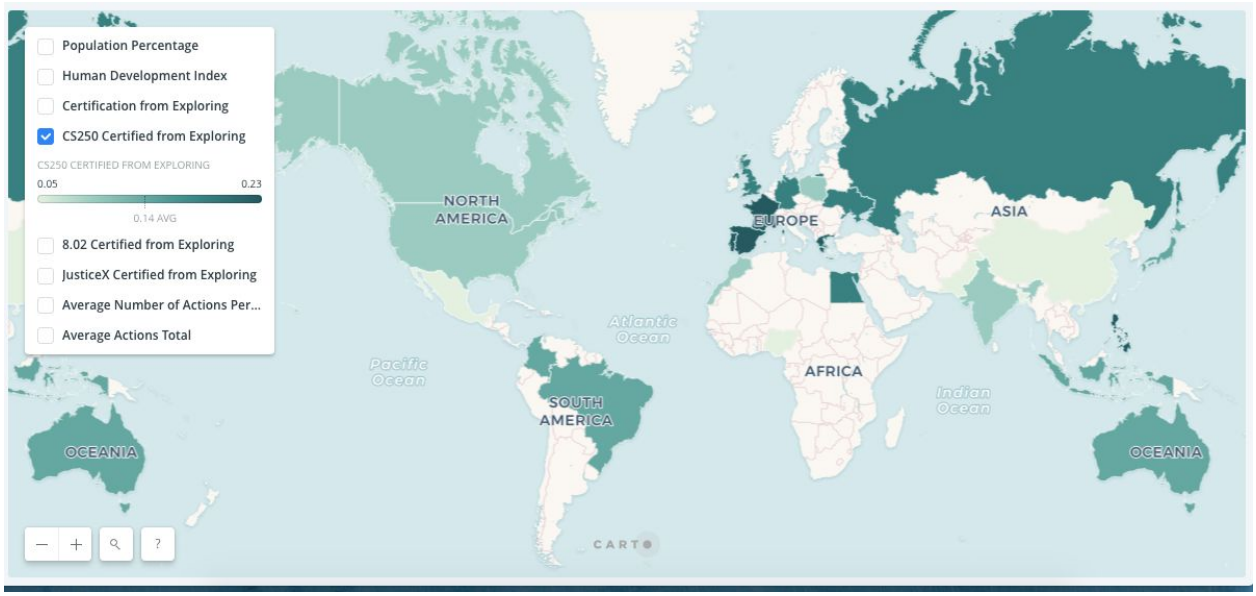

Figure 4: Rate of Certification from Exploring in CS250

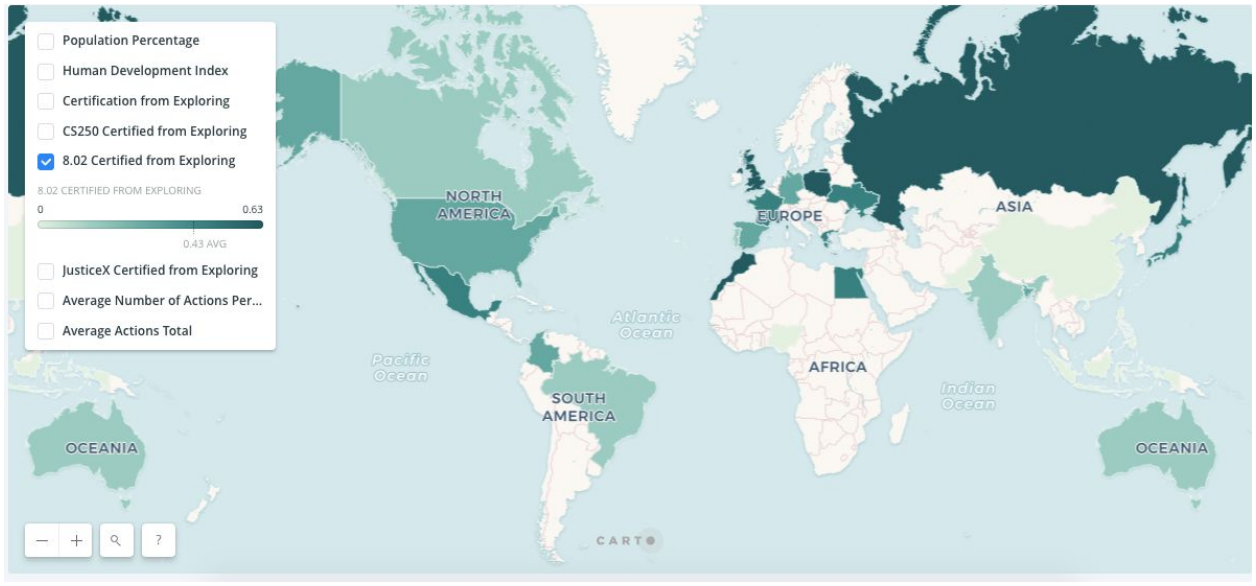

Figure 5: Rate of Certification from Exploring in 8.02

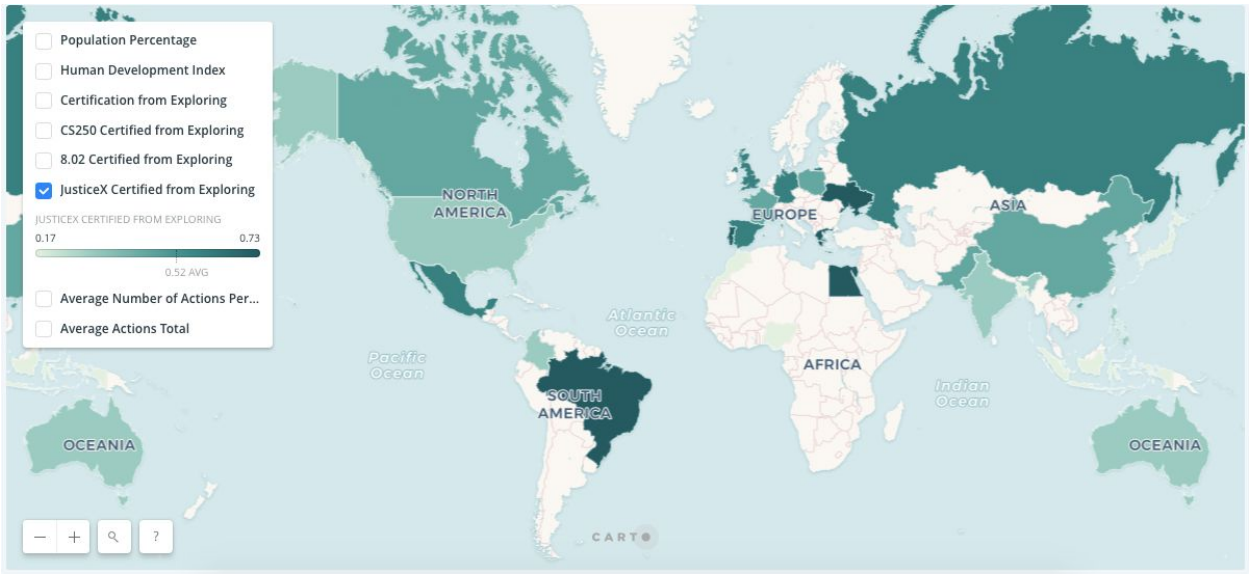

Figure 6: Rate of Certification from Exploring in JusticeX

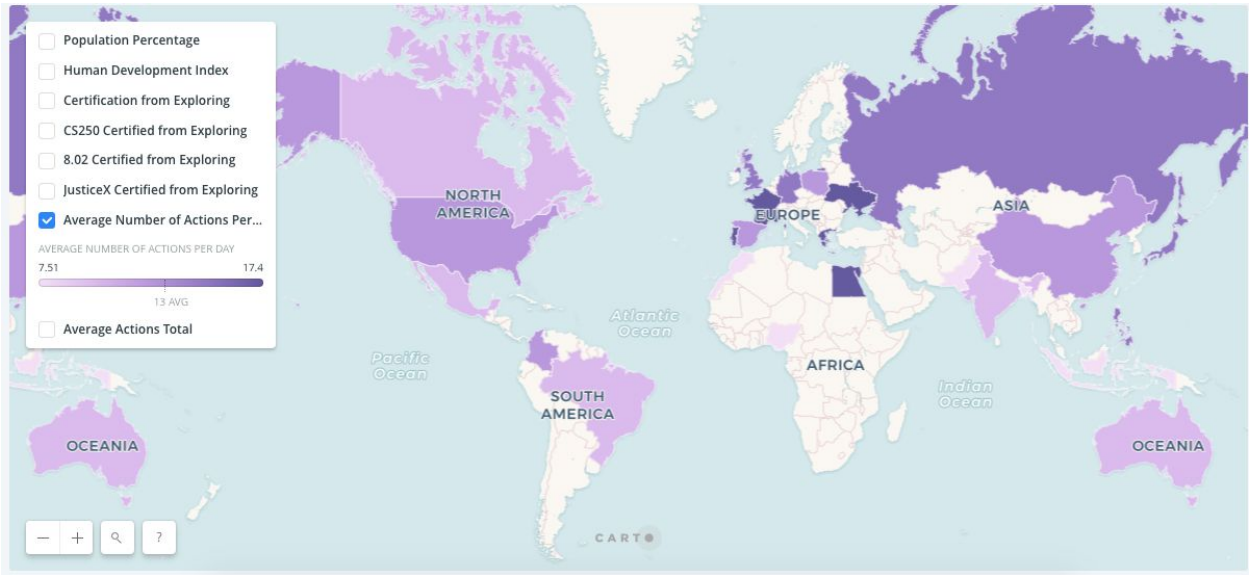

Figure 7: Average Number of Actions Per Day

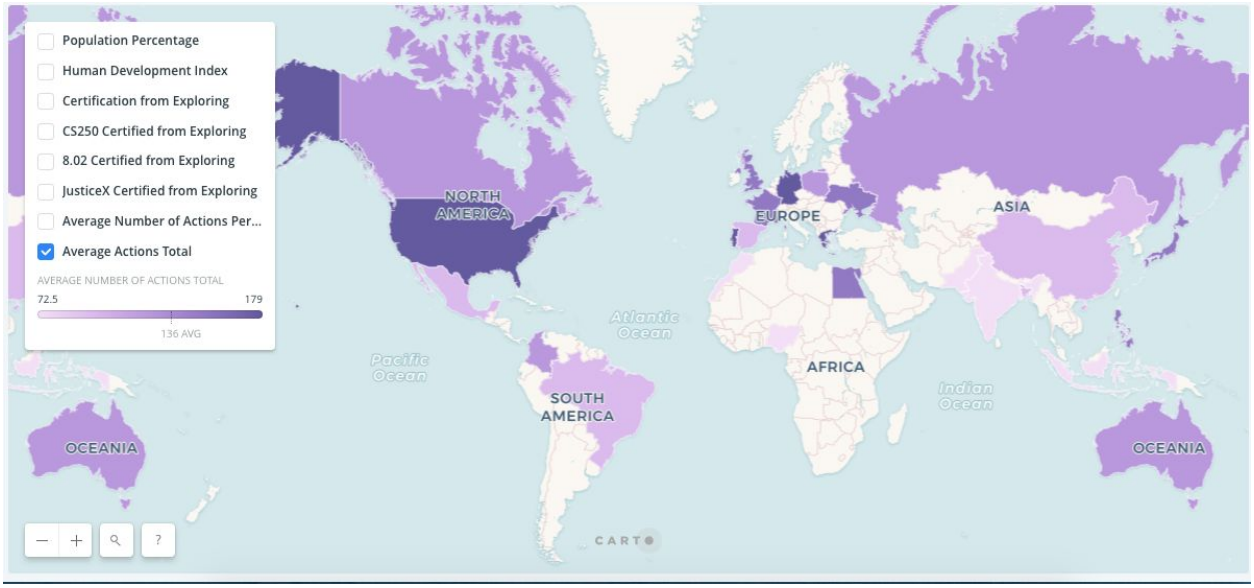

Figure 8: Average Number of Actions Total

At this stage we did not conduct quantitative analysis on our maps, but were able to qualitatively identify the following trends. There appears to be correlation between the total population percentage and human development index. North America, Australia, and Australia highin population percent and human development index. India and most countries in Africa all have lower human development index and lower population percentage, showing more correlation. It appears that people in more developed countries are more likely to be edX users. This discovery suggests that content creators at edX should be considering new ways to combat this problem and make it a more equitable resource.

In our data analysis we had some limitations which could have impacted our results. We could only map locations that were countries and not any regions such as “Other South America” or other regions that were not confined to a country. Many users choose not to disclose their country of origin, so that also eliminated a lot of data points. There are also many countries that are not included in this dataset; because of this, there are many regions of the globe that are not properly represented. Finally, the population using edX in a country not necessarily representative of its inhabitants.

We have room for plenty of further analysis. We could do quantitative analysis on the relationship between the human development index and the factors we analyzed within the dataset. So far we have only done qualitative analysis of our maps, and a quantitative analysis allow for more precise findings. We also would hope to further connect our results to other cultural factors, like the Hofstede Dimension discussed in the MOOCs paper, human development index, and other geographical data.

Reproducing your work

```
portugal = dplyr::filter(data, final_cc_cname_DI == "Portugal", explored == 1, course_id=="insert_desired_course")
philippines = dplyr::filter(data, final_cc_cname_DI == "Philippines", explored == 1, course_id=="insert_desired_course")
morocco = dplyr::filter(data, final_cc_cname_DI == "Morocco", explored == 1, course_id=="insert_desired_course")

x = list("United States", usa, "Russia", russia, "Canada", canada, "Australia", australia,
         "France", france, "Mexico", mexico, "India", india, "Japan", japan, "Colombia",
         colombia, "Germany", germany,"Poland", poland, "Indonesia", indonesia,
         "Bangladesh", bangladesh, "China", china, "United Kingdom", uk, "Ukraine", ukraine,
         "Spain", spain, "Greece", greece, "Pakistan", pakistan, "Brazil", brazil, "Nigeria",
         nigeria, "Egypt", egypt, "Portugal", portugal, "Philippines", philippines,
         "Morocco", morocco)
count = 1
for(country in x){
  if(count%%2 == 1) {
    print(country)
    }
  if(count%%2 == 0){
    print(mean(country$ndays_act, na.rm=TRUE))

  }
  count = count + 1
}
```

Here is the code we used to filter our data and gathers the statistics needed to create our heat maps. We first filtered by each country (three shown above), looking only at users who explored the course. We first analyzed these filtered datasets. For efficiency, I put each country's filtered data in a list (with the string name preceding the dataset), and then iterated through this list to print out the desired statistics of users from each country. After analyzing these filtered datasets, we proceeded to add filters by three different courses (cs50, 8.02x, and JusticeX) and then did the same iteration to print out the desired statistics again.

We then used the online mapping software called carto to plot the maps. Carto takes in CSV files and if there is some sort of geographic indicator on one of the columns allows users to color their map as such. Carto is a good platform for this level of analysis, as it is online and simple, but does not have the same analytical capabilities as other softwares like ArcGIS, which would need to be used if this project were to be expanded. Within Carto, we created the different maps by setting the style to be by value, and then creating a different map for each variable we wanted analyze.

References

Chatti, M., Dyckhoff, A., & Thüs, H. (2012). A Reference Model for Learning Analytics. *A Reference Model for Learning Analytics*. Retrieved February 24, 2019, from https://www.thues.com/upload/pdf/2012/CDST12_IJTEL.pdf.

Ferguson, Rebecca (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6) pp. 304–317. Retrieved February 24, 2019. http://oro.open.ac.uk/36374/1/IJTEL40501_Ferguson%20Jan%202013.pdf

Liu, Z., Brown, R., & Lynch, C. (2016). MOOC Learner Behaviors by Country and Culture; an Exploratory Analysis. *Educational Data Mining*. Retrieved February 25, 2019. http://www.educationaldatamining.org/EDM20x16/proceedings/paper_121.pdf

CMS.594/CMS.894 Education Technology Studio
Spring 2019