Data Visualization - Assignment 4

Vignesh Bondugula, Abhinav H Kamath, Maruthi Sriram
IMT2019092, IMT2019001, IMT2019068

Abstract

This report describes the observations, inferences and conclusions done on HarvardX–MITx Person-Course dataset. The purpose of this assignment is to implement a visual analytics workflow as a team. We built the analytics workflow and integrated all the views using a data visualization tool called Tableau.

1 Introduction

Visual analytics methodologies are useful to gleaning knowledge from data which can be used for further exploration. Because data is explored in levels, there is a concept of feedback loops. In order to build a workflow, we first make a list of analytical questions that we would like to ask of the dataset. This forms a crucial element, as any further data transformations and visual mappings are based on the list of analytical tasks. Once the list is identified, we move on to making data transformations and visual mappings which will help in building the views. In each of these steps, we can have any number of feedback loops based on our perception, in order to answer our analytical questions. Figure 1 shows the visual analytics workflow. As part of A4, we used python to clean the data a bit and then used a visual analytics tool called tableau to show the visual analytics workflow.

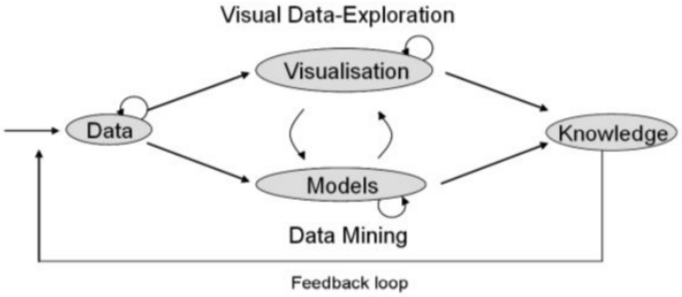


Figure 1: Visual analytics work flow. Image courtesy: Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges (pp. 154-175). Springer Berlin Heidelberg.

2 Data Overview and pre-processing

The HarvardX–MITx Person-Course dataset AY2013 contains data from first year of HarvardX and MITx courses. This dataset is at the level of one row per-person, per-course. The courses included in the dataset are as follows,

HarvardX/CB22x/2013_Spring, HarvardX/CS50x/2012, HarvardX/ER22x/2013_Spring, HarvardX/PH207x/2012_Fall, HarvardX/PH278x/2013_Spring.

The dataset contains various columns namely,

- **Course_id:** administrative, string, identifies institution (HarvardX or MITx), course name, and semester, e.g. “HarvardX/CB22x/2013_Spring”.
- **Userid_DI:** administrative, string, first portion identifies dataset (MHxPC13 corresponds to MITx HarvardX Person-Course AY13), second portion is a random ID number. Example ID: “MHxPC130442623”.
- **Registered:** administrative, 0/1; registered for course, =1 for all records in person course.
- **Viewed:** administrative, 0/1; anyone who accessed the ‘Courseware’ tab (the home of the videos, problem sets, and exams) within the edX platform for the course. Note that there exist course materials outside of the ‘Courseware’ tab, such as the Syllabus or the Discussion forums.
- **Explored:** administrative, 0/1; anyone who accessed at least half of the chapters in the courseware (chapters are the highest level on the “courseware” menu housing course content).
- **Certified:** administrative, 0/1; anyone who earned a certificate. Certificates are based on course grades, and depending on the course, the cutoff for a certificate varies from 50%-80%.
- **Final_cc_cname_DI:** mix of administrative (computed from IP address) and user provided (filled in from student address if available when IP was indeterminate); during de-identification, some country names were replaced with the corresponding continent/region name. Examples: “Other South Asia” or “Russian Federation”.
- **LoE:** user-provided, highest level of education completed. Possible values: “Less than Secondary,” “Secondary,” “Bachelor’s,” “Master’s,” and “Doctorate.”
- **YoB:** user-provided, year of birth. Example: “1980”
- **Gender:** user-provided. Possible values: m (male), f (female) and o (other).
- **Grade:** administrative, final grade in the course, ranges from 0 to 1. Example: “0.87”.
- **Start_time_DI:** administrative, date of course registration. Example: “12/19/12”.
- **Last_event_DI:** administrative, date of last interaction with course, blank if no interactions beyond registration. Example “11/17/13”.
- **nevents:** administrative, number of interactions with the course, recorded in the tracking logs; blank if no interactions beyond registration. Example: “502”.
- **ndays_act:** administrative, number of unique days student interacted with course. Example: 16
- **nplay_video:** administrative, number of play video events within the course. Example: “52”.
- **nchapters:** administrative, number of chapters (within the Courseware) with which the student interacted. Example: “12”.
- **nforum_posts:** administrative, number of posts to the Discussion Forum. Example: “8”. roles: administrative, identifies staff and instructors, but blank as staff and instructors were removed from this release.

Pre-processing : We cleaned the dataset using pandas library of python. In the dataset we found that some of the columns have so many NULL values. We removed the rows which have more null values using the value of inconsistent flag(which shows 1 for the rows which have many null values). We removed the "roles" column from the dataset as most of the cells in the columns are blank. n_events column was also removed as it had so many NULL values in it. Many of the numerical columns like grade were represented as strings in csv. These values were converted to a floating-point number. Registrations which were done from Countries which had a value of "Unknown/Other" were removed from dataset.

3 Tasks

Based on the dataset, we decided on the aim of our project and the list of analytical tasks to solve. While our analytical tasks list might seem lofty in terms of ideas and the scale of implementation effort required, we have tried our best to achieve as much as we could in the time duration of the assignment. The aim of this project is to create dashboards which visualize the data present and analyse it and infer from the visualization. The following tasks were explored and analysed:

- Determine trends of courses' popularity among students in different countries.
- Determine when the registrations started and which time of the year the course was most accessed.
- Identify how grades of students are affected by other factors like number of chapters in the course(nchapters), number of days a student interacted with the course(ndays_act), etc.
- Compare the Grades of Students according to the level of Education.
- Identify students of which level of education accessed a course the most. to Dec-2013.

4 Visualizations

4.1 Course popularities - Dashboard 1

Firstly, we wanted to see how popular a course is among different countries. To do this task a geological map was used and the number of students which were registered in a course were plotted country wise. It was observed that many European Country students applied to the Fall courses and as the universities were present in USA, majority of people applied to courses from US itself. It was followed by India, then the European Countries. The radius of the circle is used to show the registrations done by the students from each country. The color indicates the average grade obtained by the students from that country in that course. The course-ID is used as a filter to see trends of various courses among countries. Figure 2 represents this plot.

We further wanted to see how the popularity varies among male and female, as well as among different type of domains within a country. Hence, from this visualization we created a feedback loop to take data of any particular country and further observe the trends in the course popularity. Hence, we made two more plots as shown in Figure 3 and 4.

Hence, using these visualizations we built a interactive dashboard which shows us the working of the feedback loop. The plots of registrations of courses based on gender and education change based on the selected country. Figure 5 shows us one such screenshot of the dashboard.

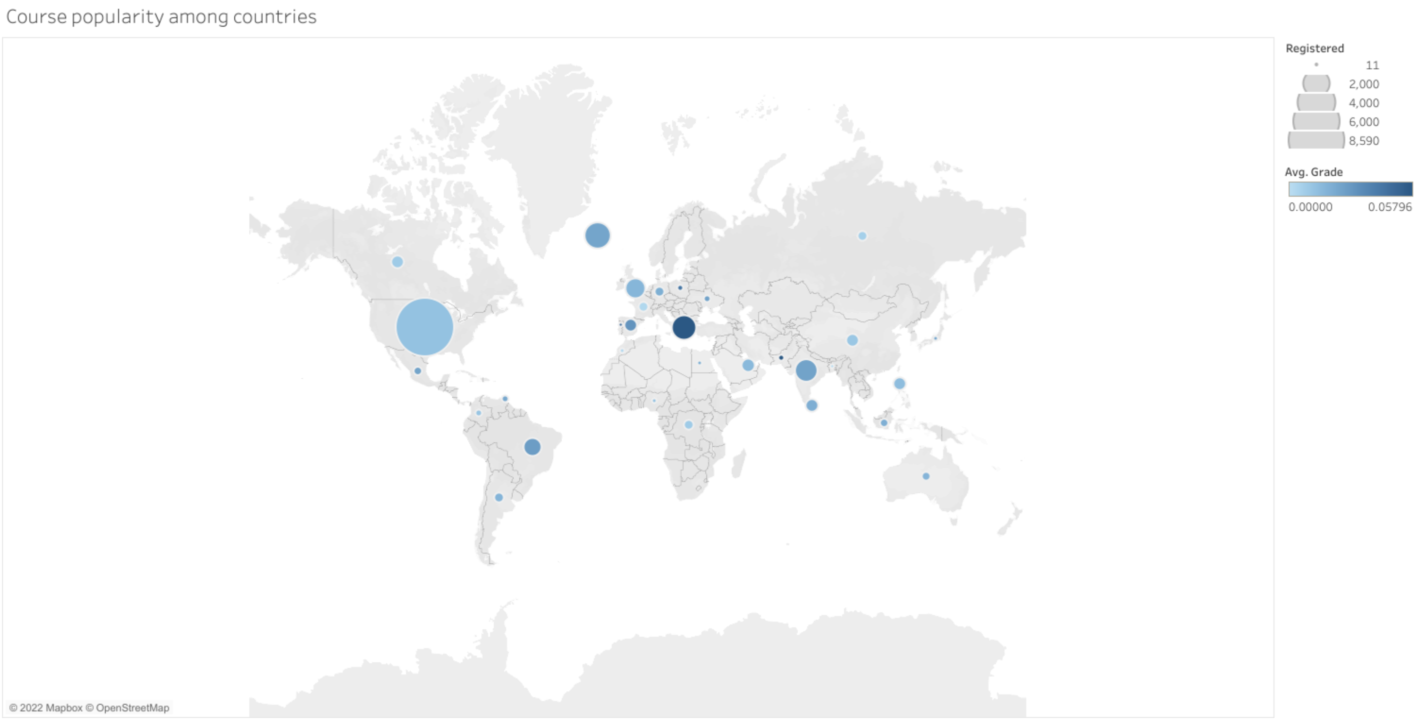

Course popularity among countries
Registered
11
2,000
4,000
6,000
8,590
Avg. Grade
0.00000
0.05796
© 2022 Mapbox © OpenStreetMap

Figure 2: The radius of the circle represents the number of registrations from the country and the color represents average grade obtained.

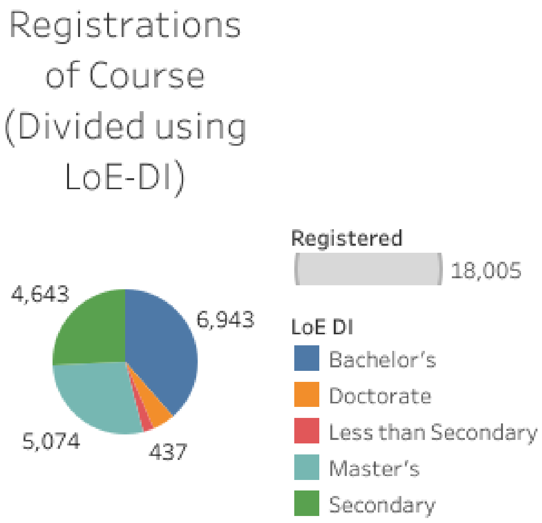


Figure 3: Registrations of a course based on different education domains.

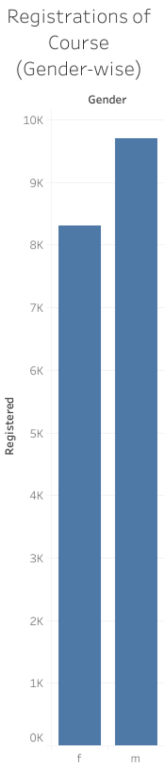


Figure 4: Registrations of a course based on gender.

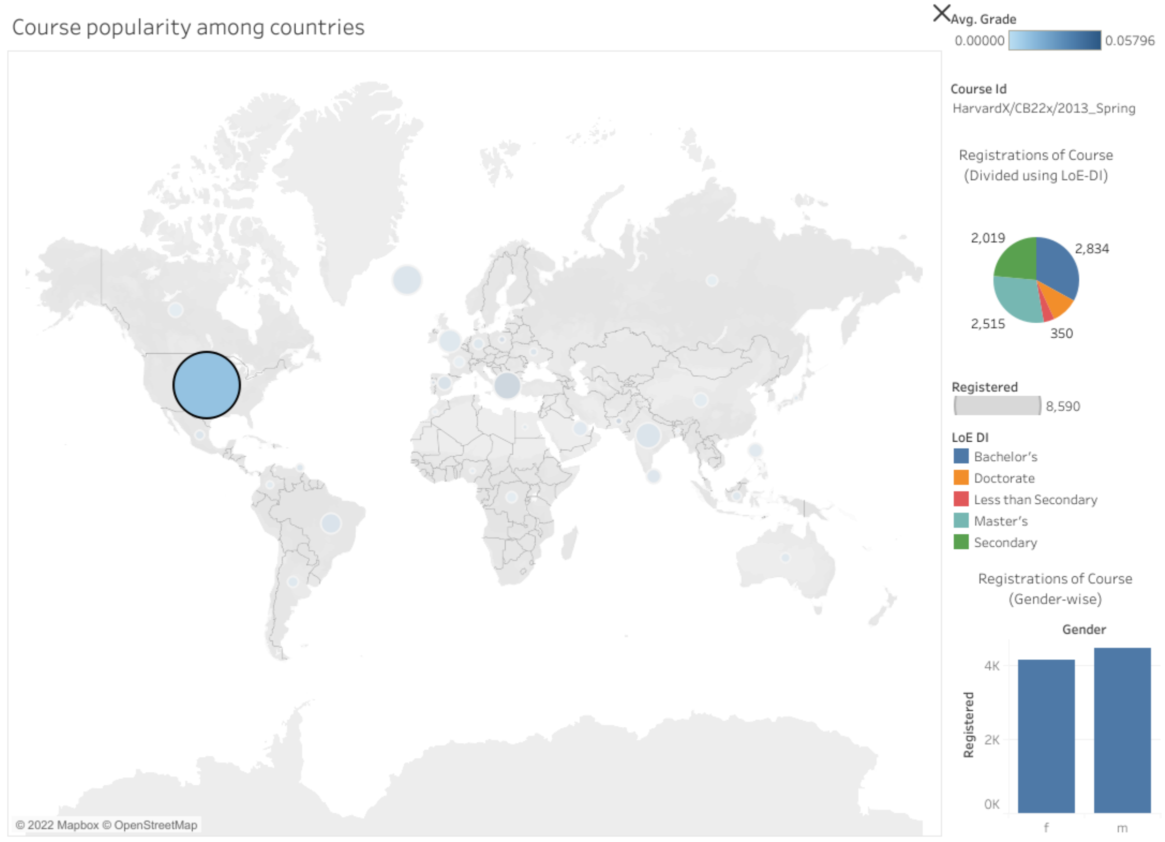


Figure 5: This shows the dashboard which visualizes course popularity at various levels (Country, Gender, Education domains).

4.2 Impact of various factors on student grades - Dashboard 2

We generated three charts which show how grade varies with respect number of days active, average number of chapters finished and average number of forum posts. Figure 6, 7 and 8 shows these plots. At first this study was conducted by seeing average grade combinedly for all courses in the same graph which was not that effective in drawing inferences. Then the graphs were made by seeing how grade varies for each course separately by seeing grade of each student in the course. This was done to see which course was giving maximum grade and what factors were effecting it.

It was observed that the number of days the student interacted with the course had the most impact on the grade of students. The course HarvardX/CS50x/2012 was the most varying one as in that course either you get a grade 0 or grade 1 by interacting the same number of days. This course also had no forum posts. For all the courses a common observation was that as you read more number of chapters, the grade you get increased linearly. The number of forum posts was not that influential on the grade of students. The n days act was influential on the grade of the course, as the grade increased as n days act value also increased. Out of all the 3 factors mentioned, Average number of chapters read by the student was the most helpful to determine the grade.

Now, we wanted to see the activity of students LoE_DI wise and grade they got and see which was the most scoring course for each category of students. For this we had to implement a dashboard with a feedback loop from the domain wise plot and the other three plots. Figure 9 shows the dashboard which is made with interactions and a feedback loop.

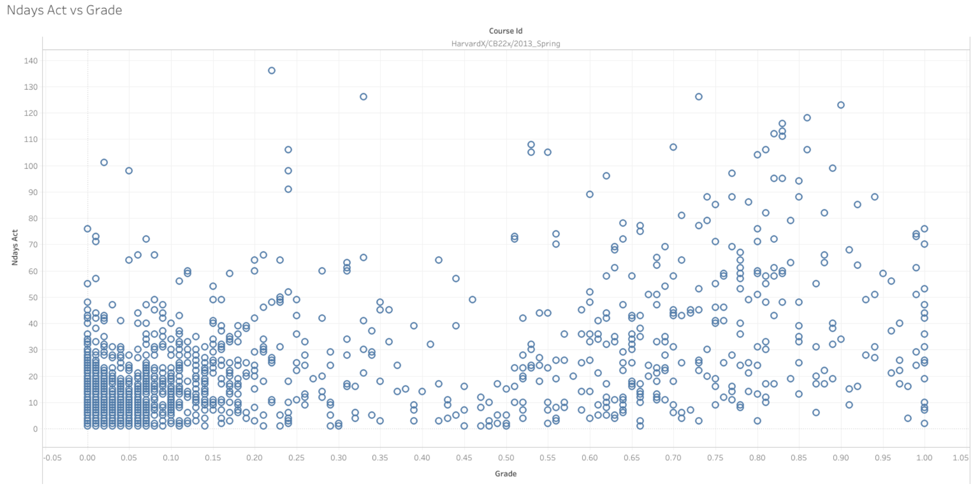


Figure 6: Scatterplot to see how number of active days effect the grades of students.

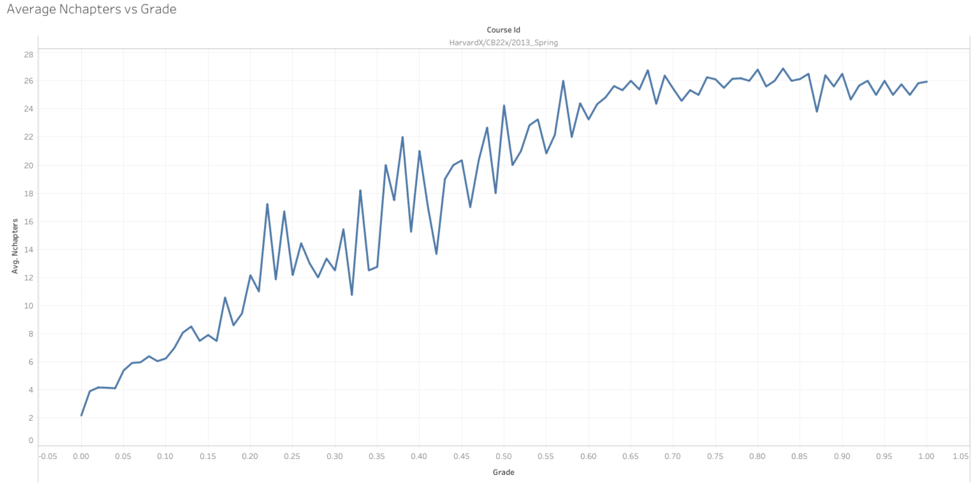


Figure 7: Line chart between average number of chapters finished and grade.

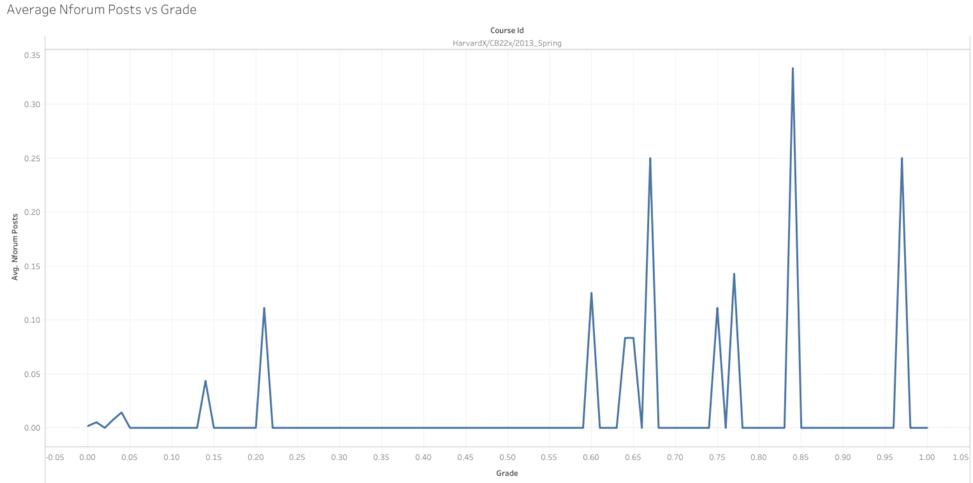


Figure 8: Line chart between average number of interactions on forum and grade.

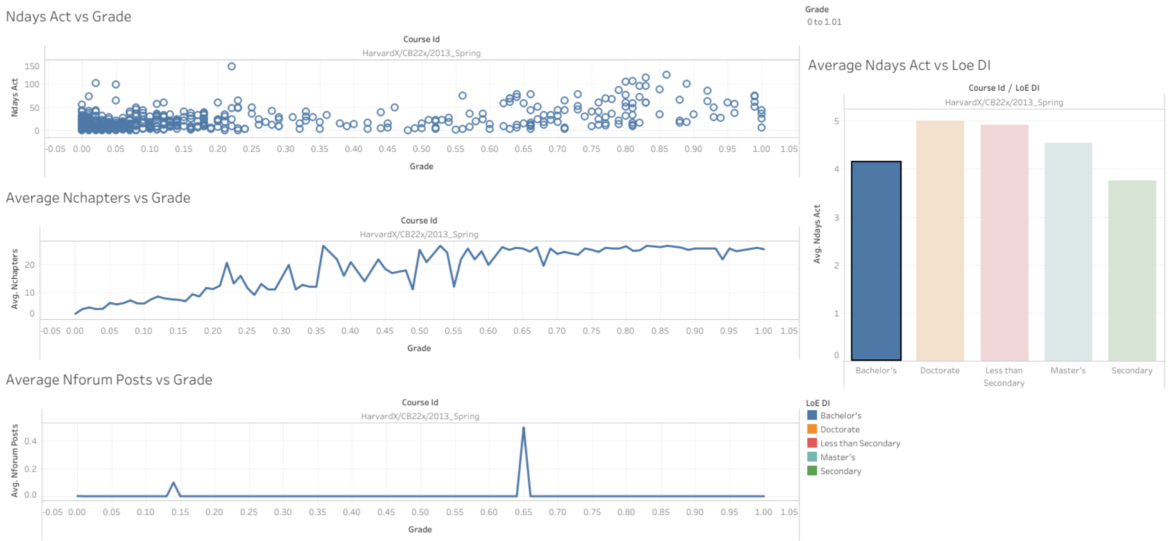


Figure 9: This figure shows the dashboard which visualizes various factors effecting grades of students in different education domains.

It was observed that "Less than Secondary" interacted with the most with all the courses except for the course HarvardX/PH207x/2012_Fall. This course was most dominated by Master's students. It was seen that HarvardX/ER22x/2013_Spring course was the toughest courses for students of all domains as the grade was not that high for all the students. The bar graph which is shown to the right was used as filter to see the grades of students domain wise. It was observed that the Doctorate people struggled in the course, HarvardX/CB22x/2013_Spring. For Bachelor's HarvardX/ER22x/2013_Spring course was more scoring, as that course was taken less by the other domain students.

5 Registrations throughout the year - Dashboard 3

This dashboard helps to see the trends of registrations of students' course wise throughout the year. We wanted to see how the registrations came in when the courses were launched. To do this task, bar graphs were plotted for each month as a bin and number of certified registrations in a month were compared with the total number of registrations. It was observed that the fall courses had more registrations on the later part of the year and the spring courses had more registrations on the starting part of the year. Figure 10 shows the registrations in every month as well as number of certifications. Since, number of certifications were very low compared to number of registrations we had to apply logarithm on number of students. The bar represents the registrations whereas the line represents number of certifcations. The average grade is also visualized per month for the certified students using a color map.

Now we also wanted to view how the registrations came in at a date level. Calendar visualization was used to plot the registrations done each day by students. The reason we did this is to understand in detail when exactly the course registrations started. We wanted to know how many days after the launch of the course the registrations started. Figure 11 shows the calendar view of the registrations.

We then created a dashboard which takes a input from the bar chart built month wise, applys a feedback loop, and displays the calendar for that month. Figure 12 shows the view of dashboard. The bar graph which is shown to the right is used as a filter for showing registrations Day wise. As you select a bar in the graph, that month's registrations are

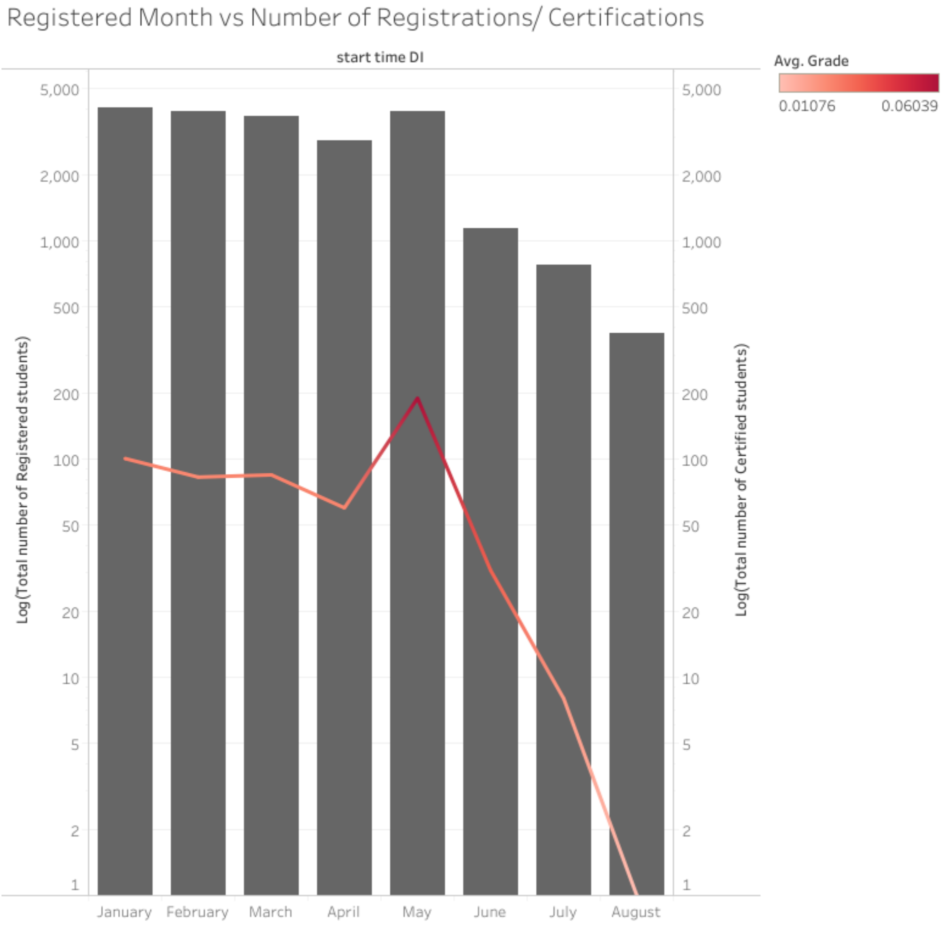


Figure 10: Number of registrations and certifications per month is visualized in the image.

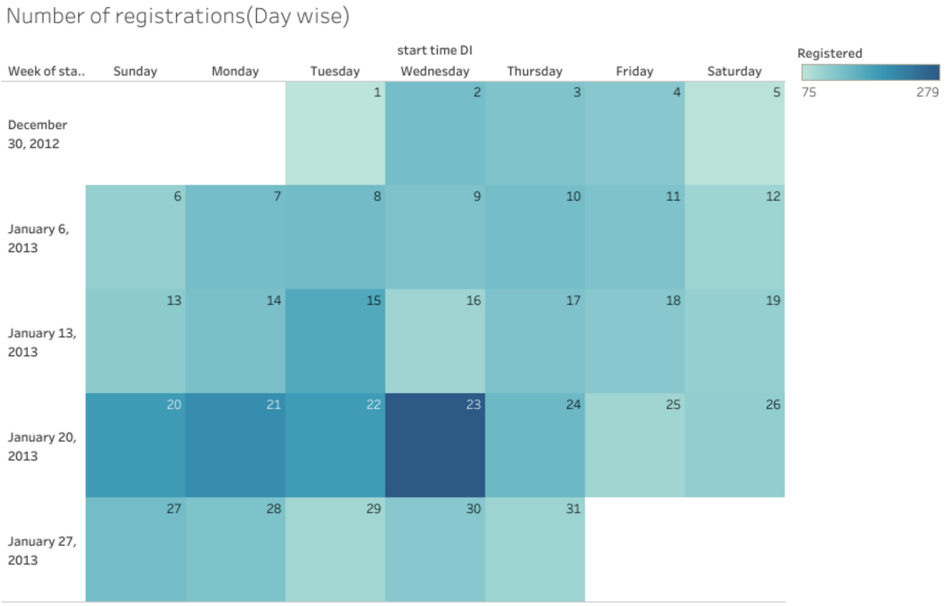
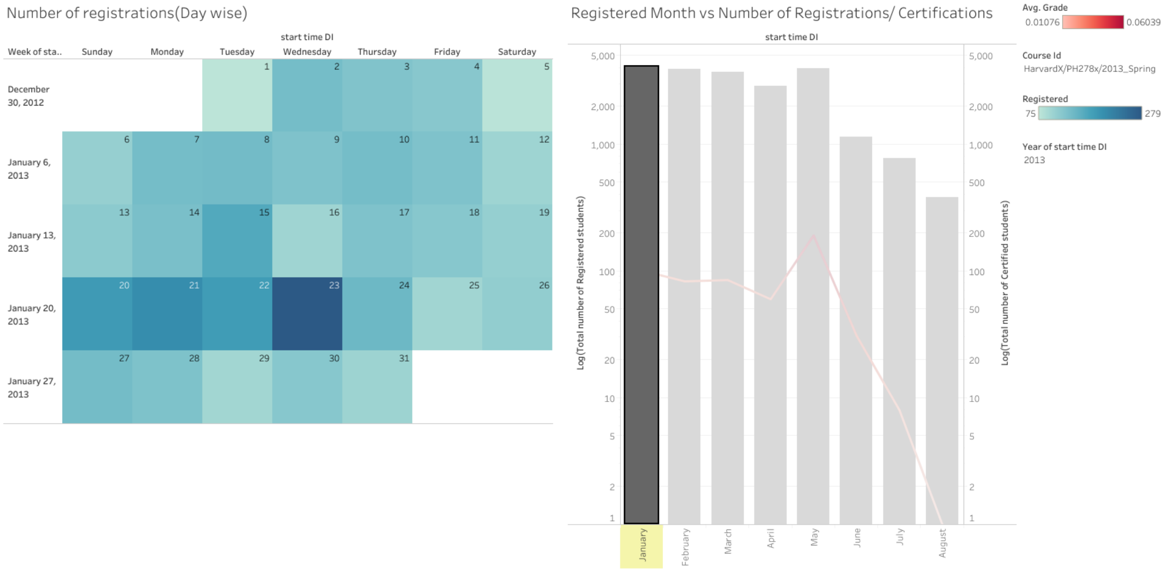


Figure 11: Number of registrations every day is visualized in the image using a Calendar visualization.

shown day wise in the Calendar. We can chose a course and year in the dashboard using the dropdown menus.



Course Id
HarvardX/PH278x/2013_Spring

Year of start time DI
2013

Figure 12: This figure shows the dashboard which visualizes day wise registrations for a particular course in a year.

6 Conclusion

In this assignment, we learned about how to design a Visual Analytics Workflow. The 2 key features of the Visual Analytics flow, namely the visualizations and the feedback loops have been studied and included in our work. The usage of tools like Tableau have been explored for doing the assignment. The visualizations gave a lot of insights.

7 Contributions

- **Vignesh Bondugula :** Made visualisations related to geography and visualised them in Dashboard 1.
- **Abhinav H Kamath :** Made visualisations related to factors effecting grades and visualised them in Dashboard 2.
- **Maruthi Sriram Rachapudi :** Made visualisations related to day/time and visualised them in Dashboard 3.

8 Files

- Certified vs Registered.twbx: This file was to see trend of registrations for course in 2013 year.
- Dashboard_12.twbx: This file contains plots and dashboard related to first two dashboards.