DATA INTENSIVE COMPUTING

PROJECT 2 STAGE 3

OSHIN SANJAY PATWA

UB Id: oshinsan

Person #: 50169203

PRITHVI GOLLU INDRAKUMAR

UB Id: pgolluin

Person #: 50169089

Department of Computer Science

State University of New York at Buffalo

INDEX

Problem statement:

Analyzing the data and performing exploratory data analysis on the data set related to State University of New York at Buffalo. This data set consists of data for related to the number of students registered, courses offered, time scheduling of classes in various hall rooms in UB campus and their maximum capacities.

Question:

Finding important characteristics and trends of UB data base about class scheduling, in order to draw some firm conclusions that can help make better class rooms scheduling decisions in regards with its utilization at UB for the future years.

Goal:

Our main goal is to understand this data and visualize the output obtained from the 9 MR algorithms. These algorithms provide solutions to list of 9 relevant questions designed by us. This helps us to realize important trends in this given data set and to extract "knowledge" from this data and the solutions of the MR Jobs.

Data Characteristics:

We referred to bina_classschedule.csv file provided by Professor Bina Ramamurthy. We worked on semi structured data in CSV format.

We designed a set of **9 complex questions each with 2 or more mappers** in stage 2 of this project and came up with map reduce solutions to these problems in an attempt to give optimized solutions to these questions. This helped us comprehend the extensive data set in a constructive way which reflected some important observations, thus helping to make better decisions in the future. This would surely help in the effective utilization of the classroom based on the courses offered in them, students registered etc.

In stage 3 of this project we have carried out extensive analysis of data on the above mentioned data set by ingesting the results of the MR analysis in tableau to create visual insights into the data. We have realized data visualization and visual tendering using tableau. The output of our MR

programs is transformed into visually appealing dashboards using tableau. We have framed a separate dashboard for a set of linked question that connect our results and we have integrated these dashboards to form a story that explains an overall analysis of the MR jobs implemented and data set visualized.

Implementation details were discussed in detail in the earlier report i.e. problem 2 stage 2, in this part of the report we have focused on the analysis part and data visualization part of the results obtained from implementing Map Reduce jobs.

Data Analysis:

Following is the detailed analysis of various dashboards designed, and an interpretation of individual worksheets in it.

Dashboard 1 Theme: Top 20 courses for which maximum students have registered and trends of the course for which maximum students have registered over every year semester.

MR Code Implementation: This dashboard infers results obtained from 5 different MR Jobs of which first 3 MR constitute of first question: Top 20 courses for which maximum students have registered (Problem 1) and the last 2 MR Jobs constitute the seconds question: Which year semester had the most number of students registered for General Chemistry and also their trends each year semester (Problem 2).

Job 1: Found the number of students registered for each course.

Job 2: To sort the records on values.

Job 3: To display top 20 records.

Job 4: Found the number of students registered for General chemistry for each year semester respectively.

Job 5: To sort the records on values.

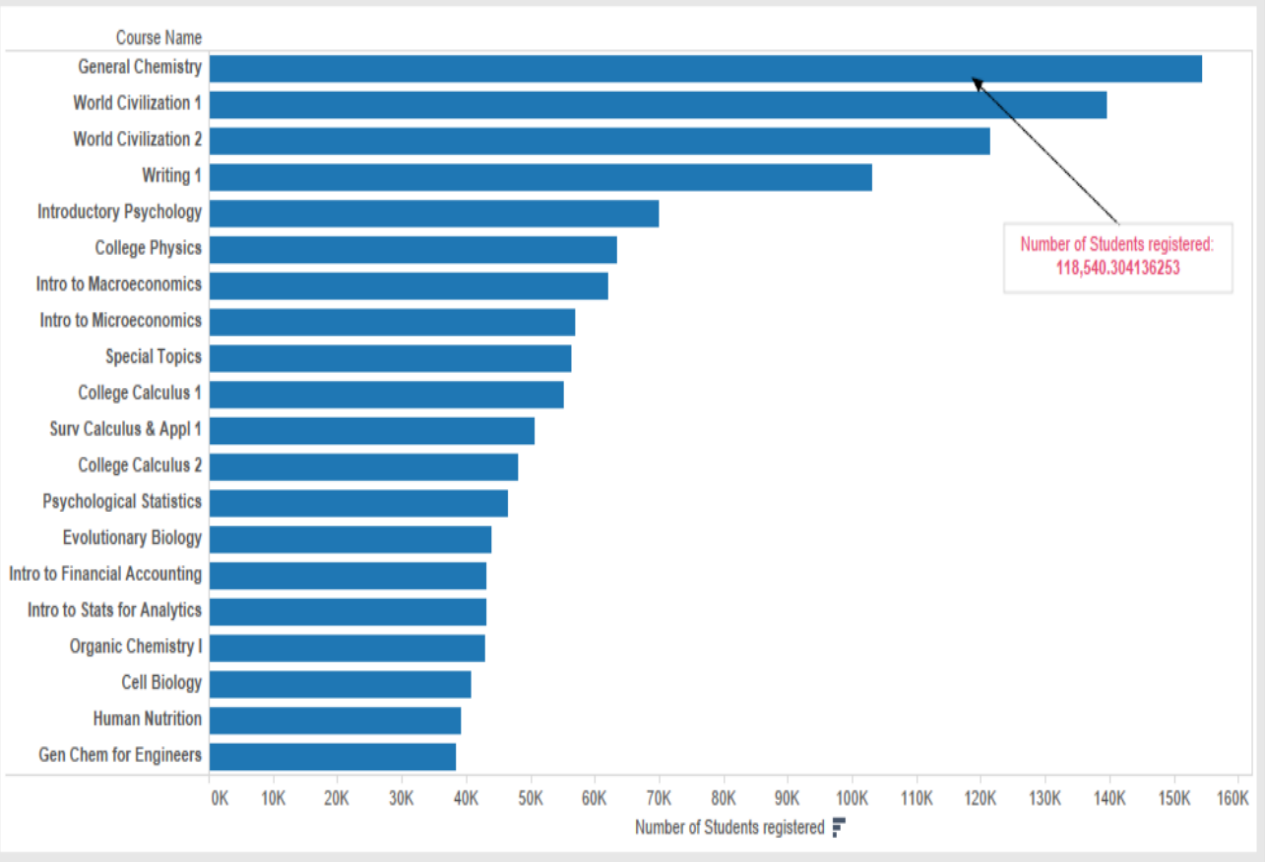


Fig 1: Top 20 courses for which maximum students registered.

From Fig1, we can see that maximum number of students have registered for General Chemistry up till now. Besides, World Civilization 1 and World Civilization 2 are also in demand. These courses are more in demand at UB according the data provided and hence they should be allocated comparatively greater lecture rooms or lecture rooms having greater capacity so as to avoid mismanagement. These courses should be allocated a greater capacity as they seem to be more demand.

From fig 2, we again see that maximum students had registered for General Chemistry last year in Fall semester i.e. Fall 2015. Also as we can observe from the output that greater number of students have registered for this subject in Fall and Spring semesters. Hence this subject should be allocated larger or greater number of classrooms in Fall and Spring semesters. The course capacity for General Chemistry should be probably increased for these semesters so that students can get the benefits of the course.

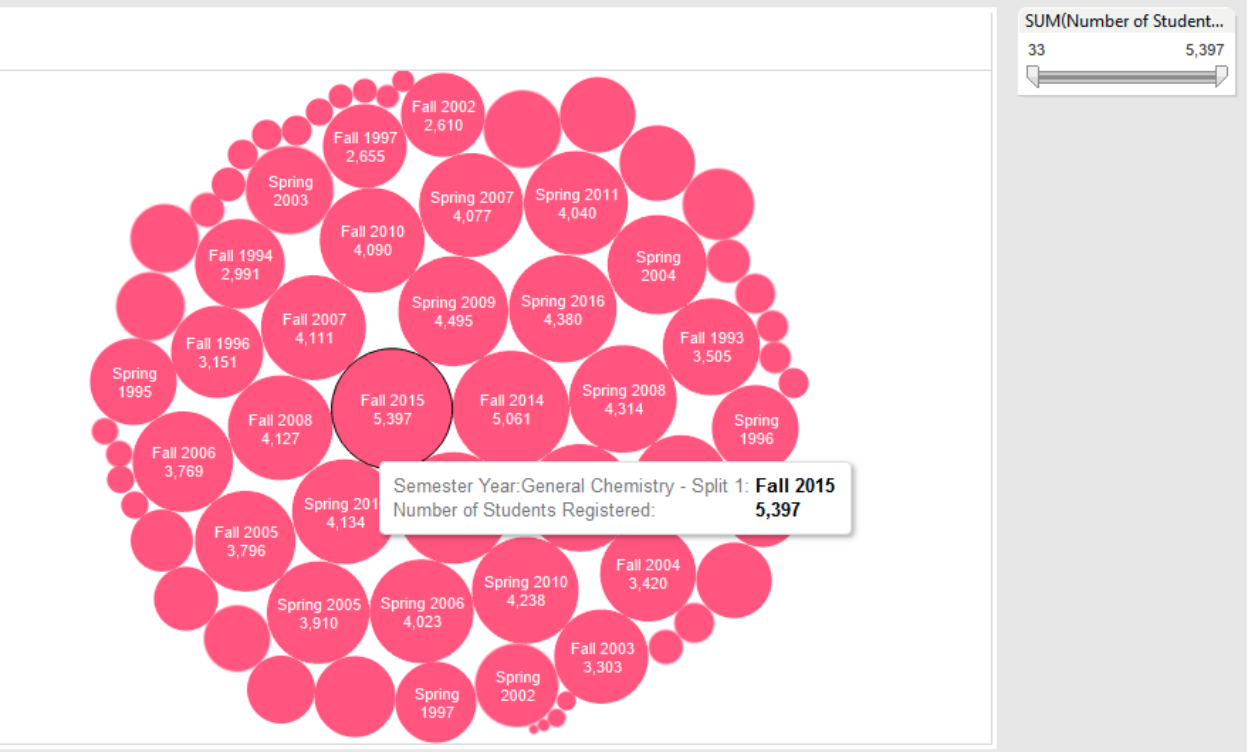


Fig 2: Students registered for General Chemistry in each year semester.

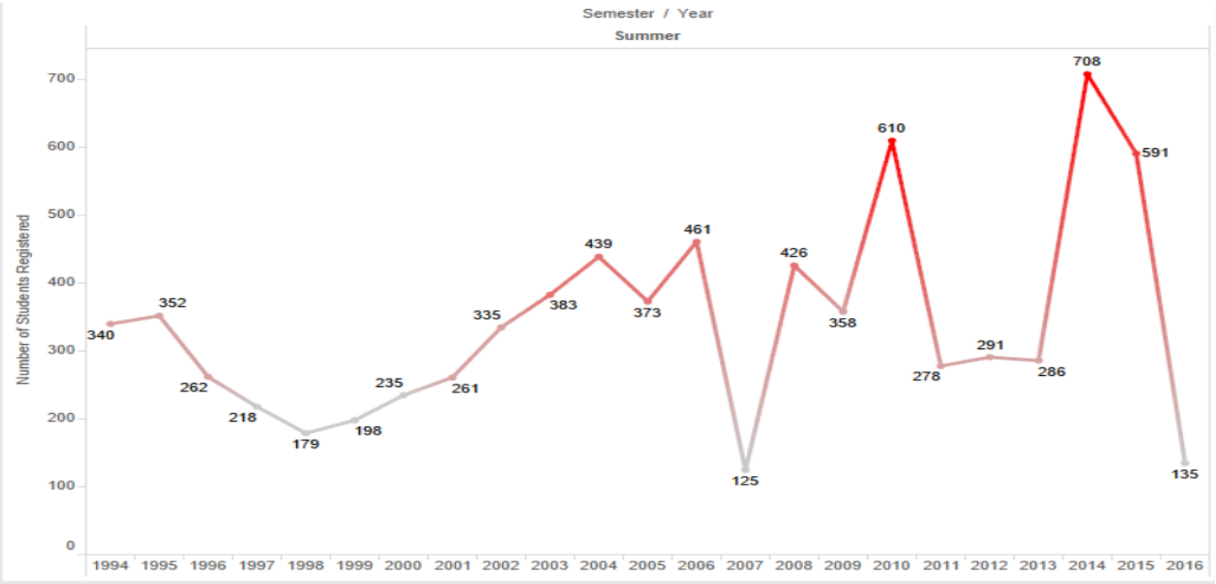


Fig 3a: Trends for students registered in general chemistry over the years for summer semester

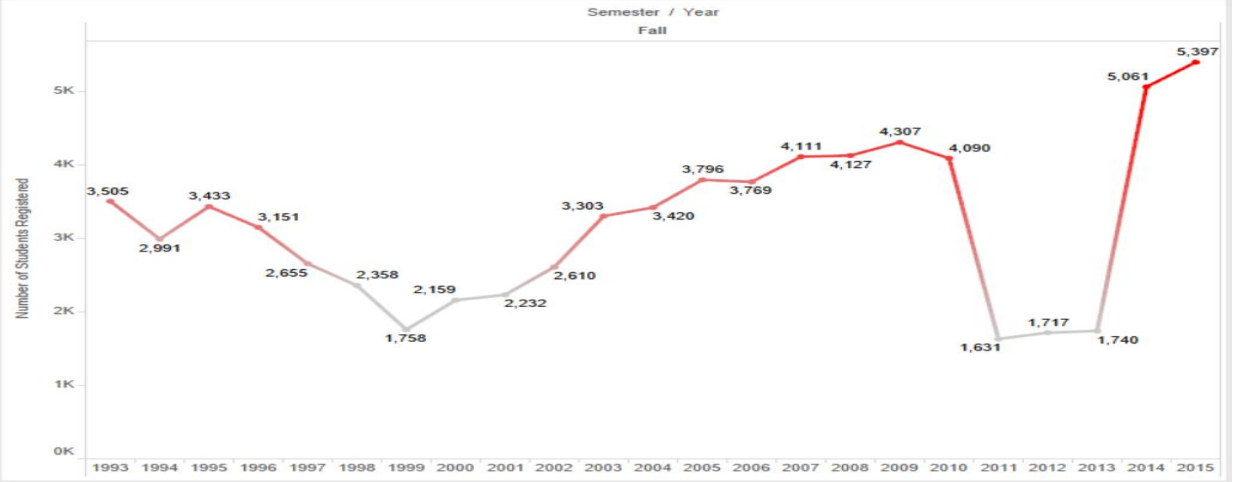


Fig3b: Trends for students registered in general chemistry over the years for Fall semester

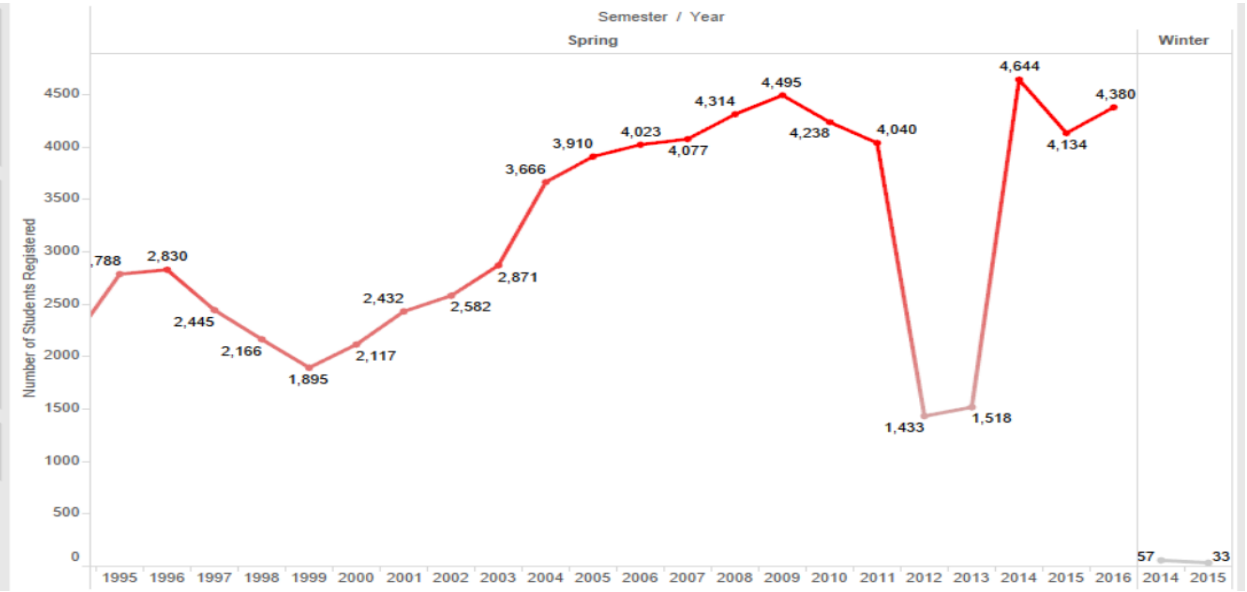
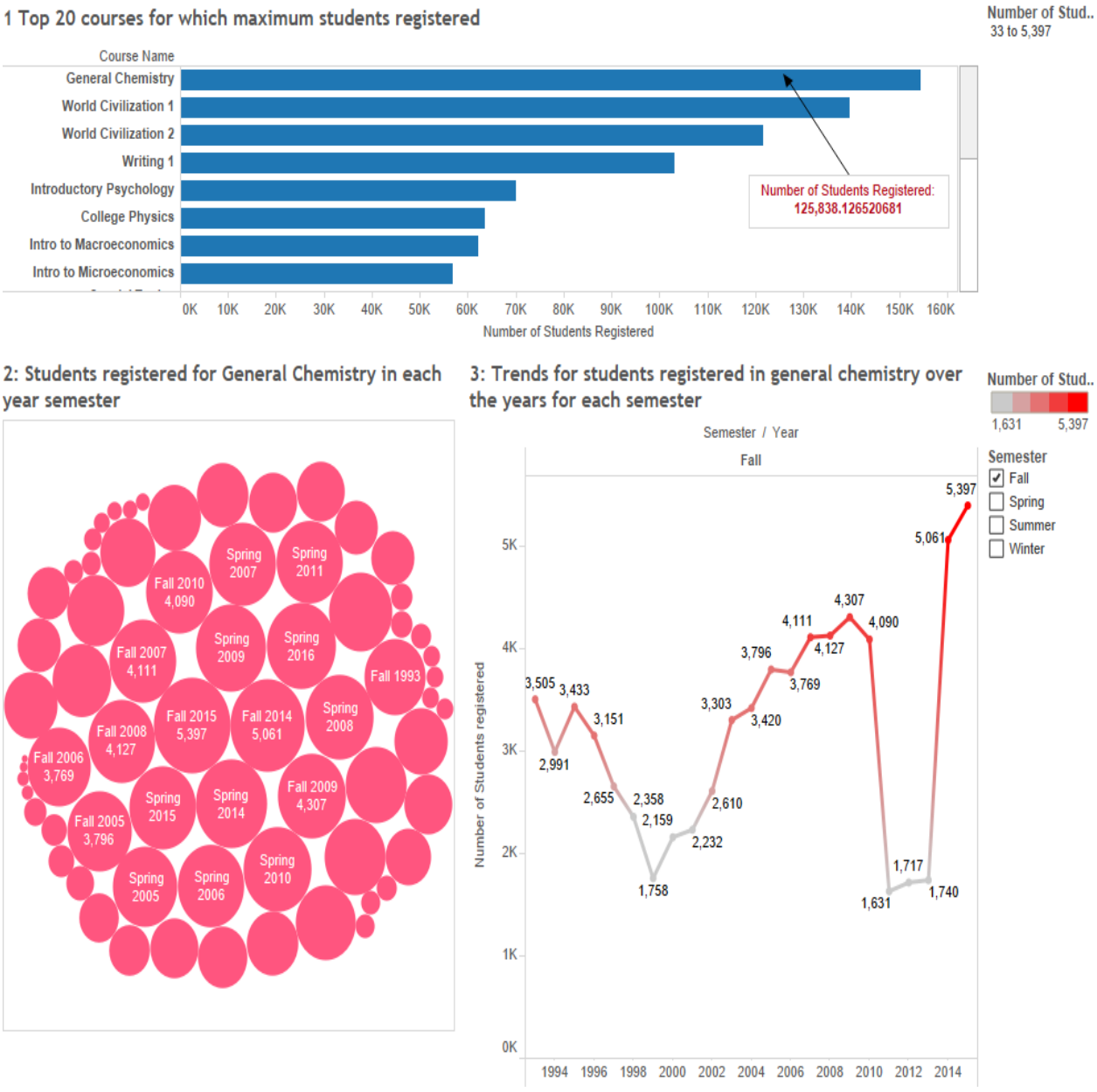


Fig 3c: Trends for students registered in general chemistry over the years for Spring semester

As seen from fig 3, we have also shown trends of students registered for General Chemistry in each year, semester wise i.e. we have shown trends of students registered for General chemistry for all years. As we can see very less students have registered for General chemistry in winter semesters so far. In summers there was a sudden fall in the students registered for General Chemistry in 2007, 2011 and 2016. However, for the Fall semesters there was never a sharp fall

in the number of students registered except in 2011 to 2013 this course seems to be in continuous demand in the fall semester. For spring semester too there was a continuous increasing demand for this course except a fall was experienced in 2011-2012. The was sudden decrease in the number of students registered for General chemistry in 2011 for spring, summer's and fall semesters as well, there may have been some situation like professor's unavailability, not sufficient facilities etc. that may have caused this sudden fall in the student's registration number.



Dashboard 1

Dashboard 2 Theme: Hall that utilized maximum time since year 2000 and trends of time utilized by that particular hall throughout the consecutive years.

MR Code Implementation: This dashboard infers results obtained from 4 different MR Jobs of which first 3 MR constitute of first question: Hall that utilized maximum time since year 2000 (Problem 3) and the last 2 MR Jobs constitute the seconds question Showing the trends for time utilized by that particular hall in consecutive years (Problem4).

Job 1: Found the time utilized by each hall.

Job 2: To sort the records on values.

Job 3: To find the time utilized by Baldy hall in each year.

Job 4: To find the change in time utilization for consecutive years.

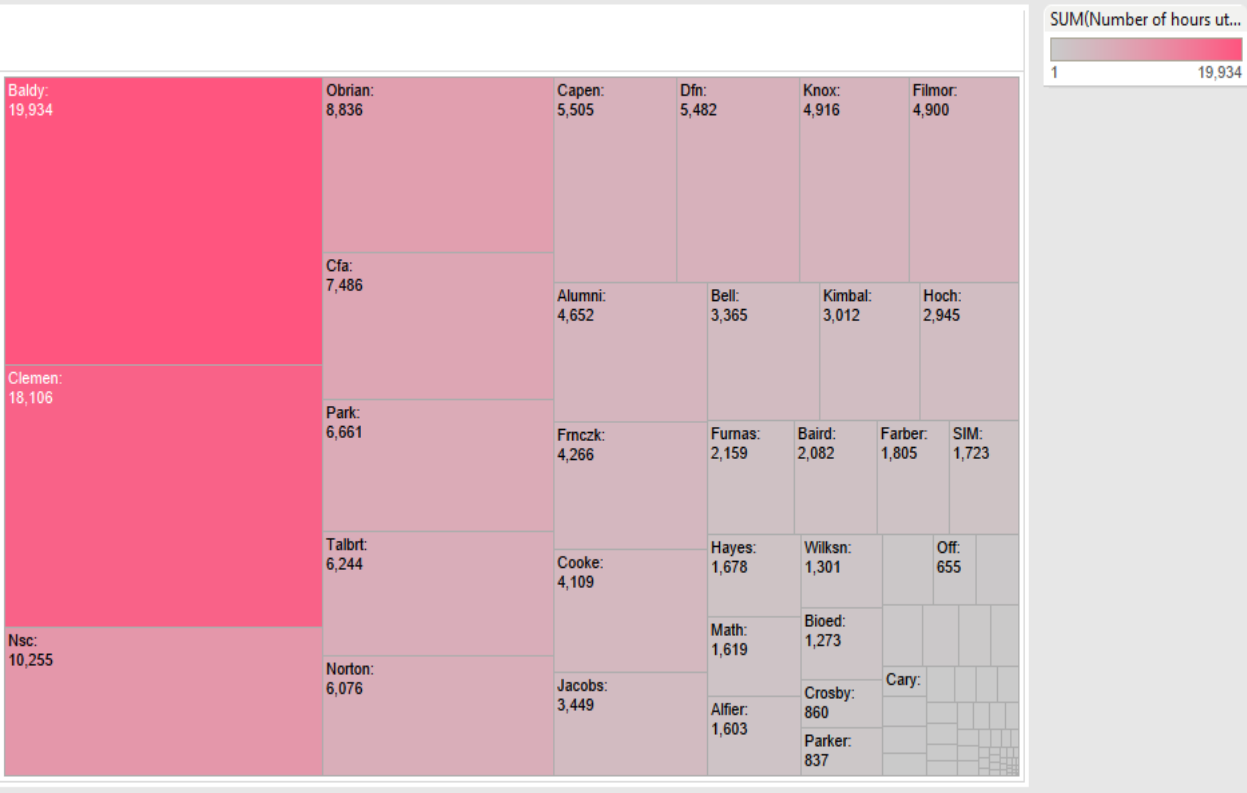


Fig 4: Time utilized by each hall since the year 2000

From fig 4, we can see that Baldy hall seems to be more occupied based on time as compared to other halls as it consumes the greatest area in the below tree map graph. The schedule for this hall should be decided well in advance as the hall seems to be occupied. Clemen, Obrian, Nsc also seem to have utilized more amount of time.

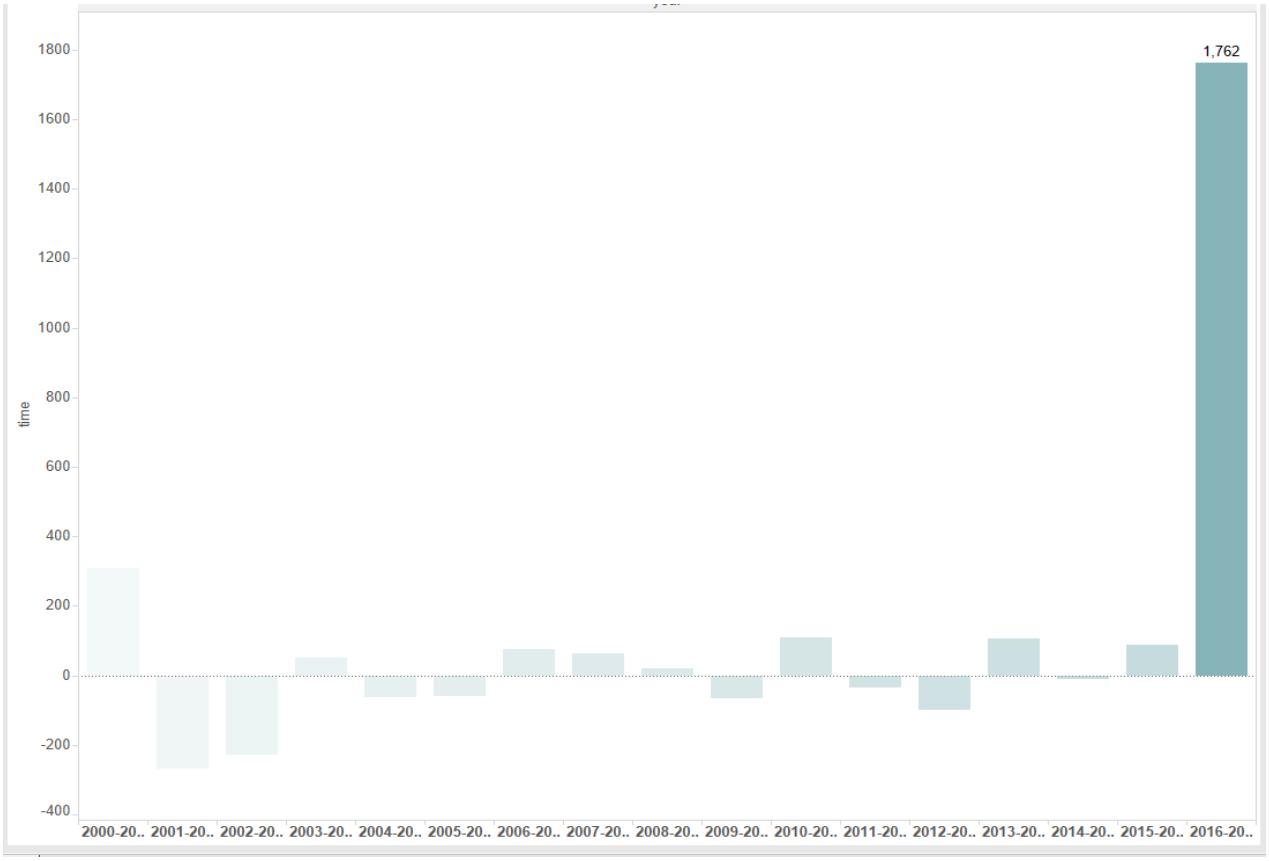
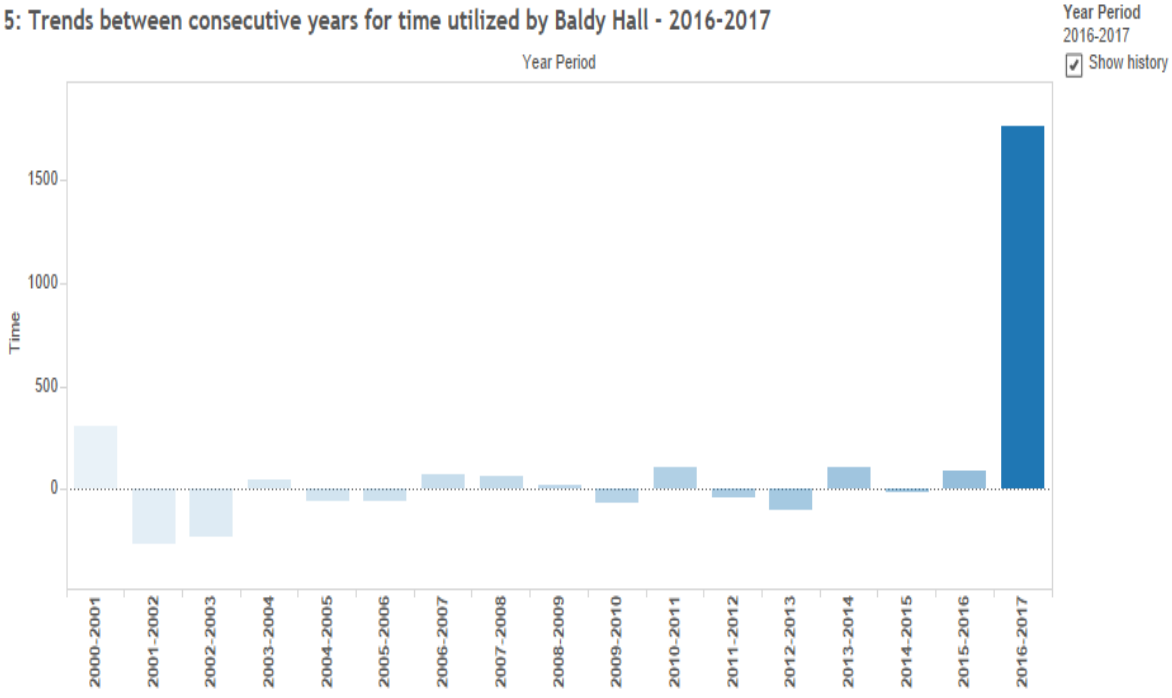


Fig 5: Trends between consecutive years for time utilized by Baldy Hall

From fig 5 we can see that there was a remarkable increase in utilization of Baldy hall from year 2016 to 2017 by 1762 hours. Roughly the time utilized by Baldy has increased over the years. Even if there was a decrease in the time utilization that was comparatively less.


4: Time utilized by each hall since the year 2000
Number of hours utilized
1
19,934
Baldy: 19,934
Nsc: 10,255
Talbrt: 6,244
Knox: 4,916
Filmor: 4,900
Alumni: 4,652
Frnczk: 4,266
Norton: 6,076
Obrian: 8,836
Cooke: 4,109
Hoch: 2,945
Furnas: 2,159
Baird: 2,082
Clemen: 18,106
Capen: 5,505
Jacobs: 3,449
SIM: 1,723
Bioed: 1,273
Cfa: 7,486
Hayes: 1,678
Bell: 3,365
Wende: 779
Dfn: 5,482
Math: 1,619
Off:
Park: 6,661
Kimbal: 3,012
Alfier: 1,603
Dfn_A:
Squire:


5: Trends between consecutive years for time utilized by Baldy Hall - 2016-2017
Year Period
2016-2017
Show history
Year Period
Time
1500
1000
500
0
2000-2001
2001-2002
2002-2003
2003-2004
2004-2005
2005-2006
2006-2007
2007-2008
2008-2009
2009-2010
2010-2011
2011-2012
2012-2013
2013-2014
2014-2015
2015-2016
2016-2017

Dashboard 2

Dashboard 3 Theme: Peak time for which most number of classes were held for each year.

MR Code Implementation: This dashboard infers results obtained from 2 different MR Jobs which basically find the peak time for which most number of classes were held each year (Problem 5).

Job 1: To get how many classes are held respective to each time period.

Job 2: To get the time for which maximum classes are held in each year respectively.

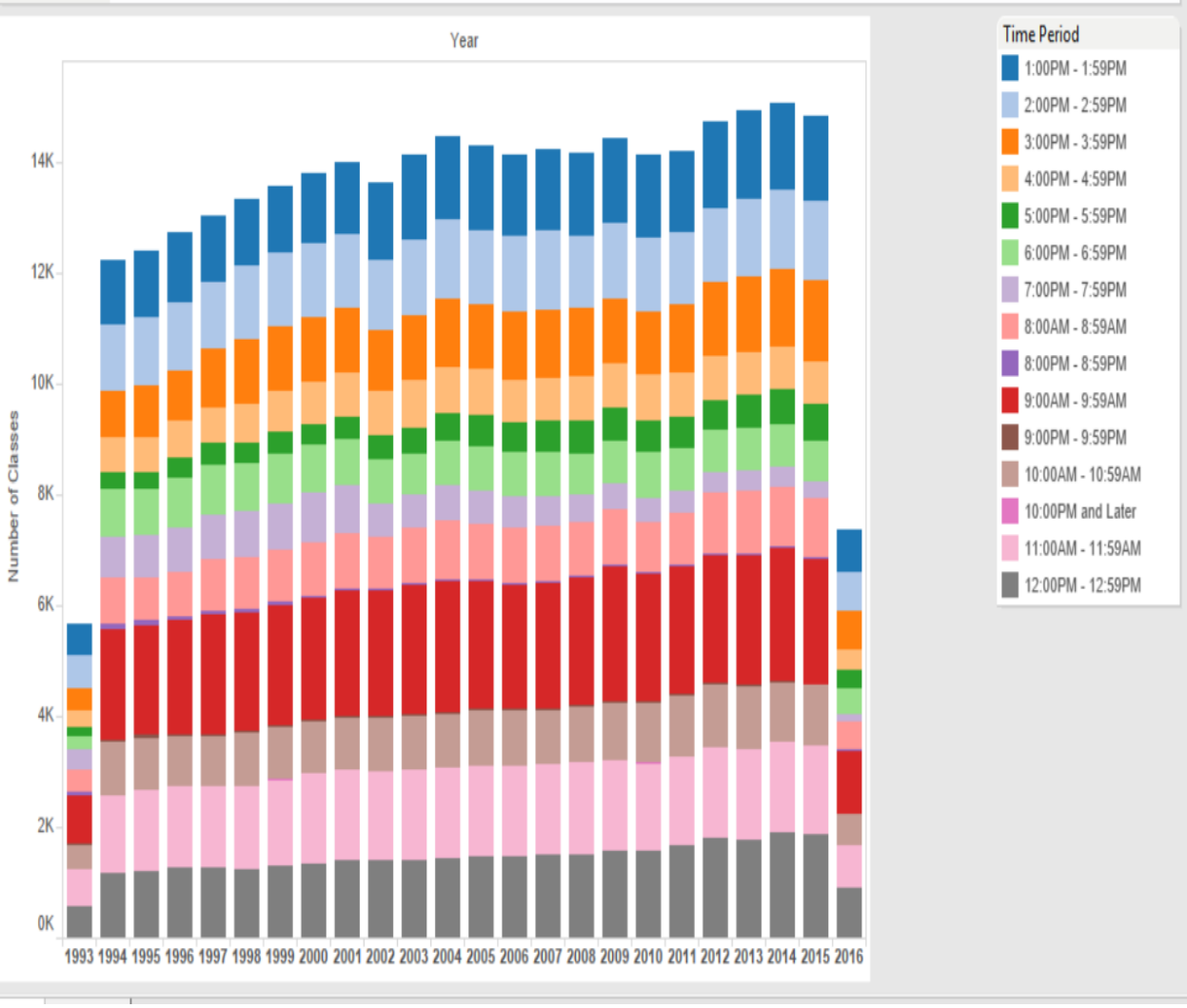


Fig 6: Number of classes that were held respective to each time period.

From fig 6, we can infer the number of classes that were held respective to each time period for each year, say in year 2014 roughly there were 2000 classes held in the time slot between 12:00 PM – 12:59 PM. From the figure given above we have also seen that the red bars have the most length in each year showing that the peak time for almost all years is 9:00AM – 9:59 AM. This slot of the day seems to be very busy in all years and there seem to be many classes going on during this time period. Over all the classrooms in UB campus will be filled with students during this period as this is the peak lecture period. Probably more class rooms should be allocated to this time period to avoid mismanagement.

The Fig 7 below shows the distribution of classes held during this time period for all years. As we can see maximum number of classes were held in this time slot in the year 2009 i.e. almost 2500 classes. Also we can see that as the years progressed there was an increase in the number of classes held during the same time slot roughly.

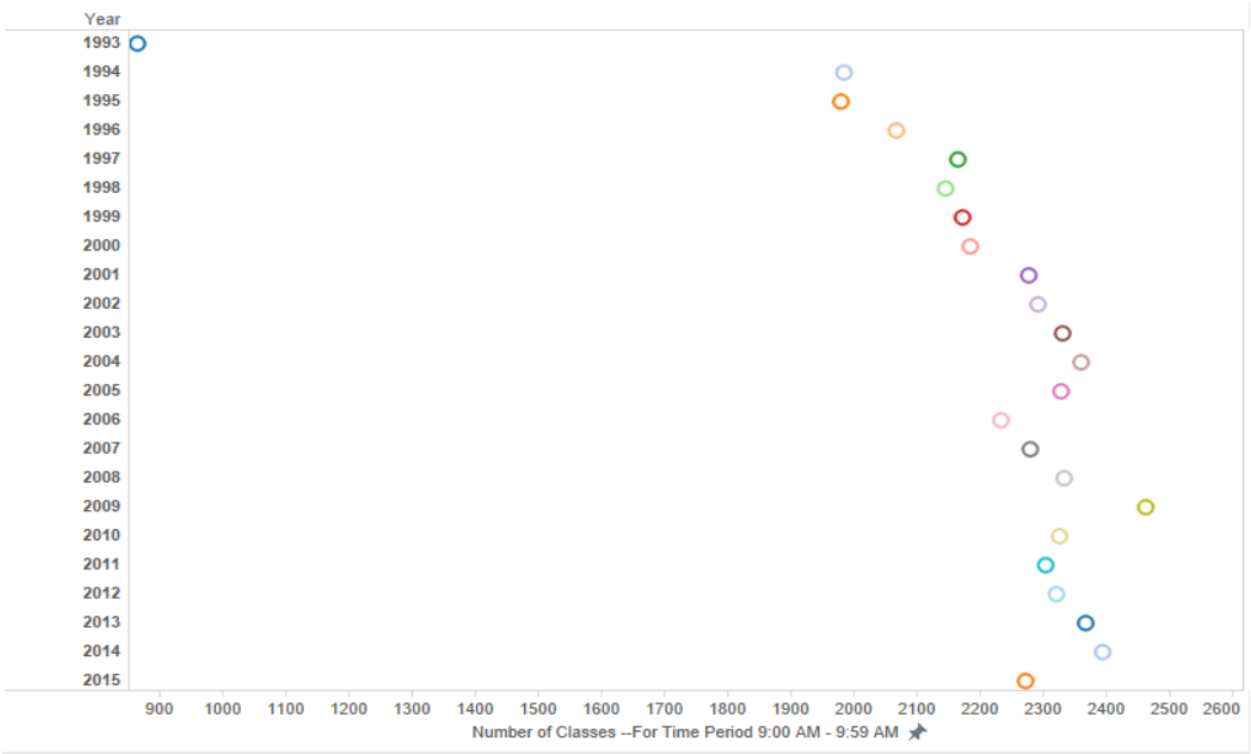
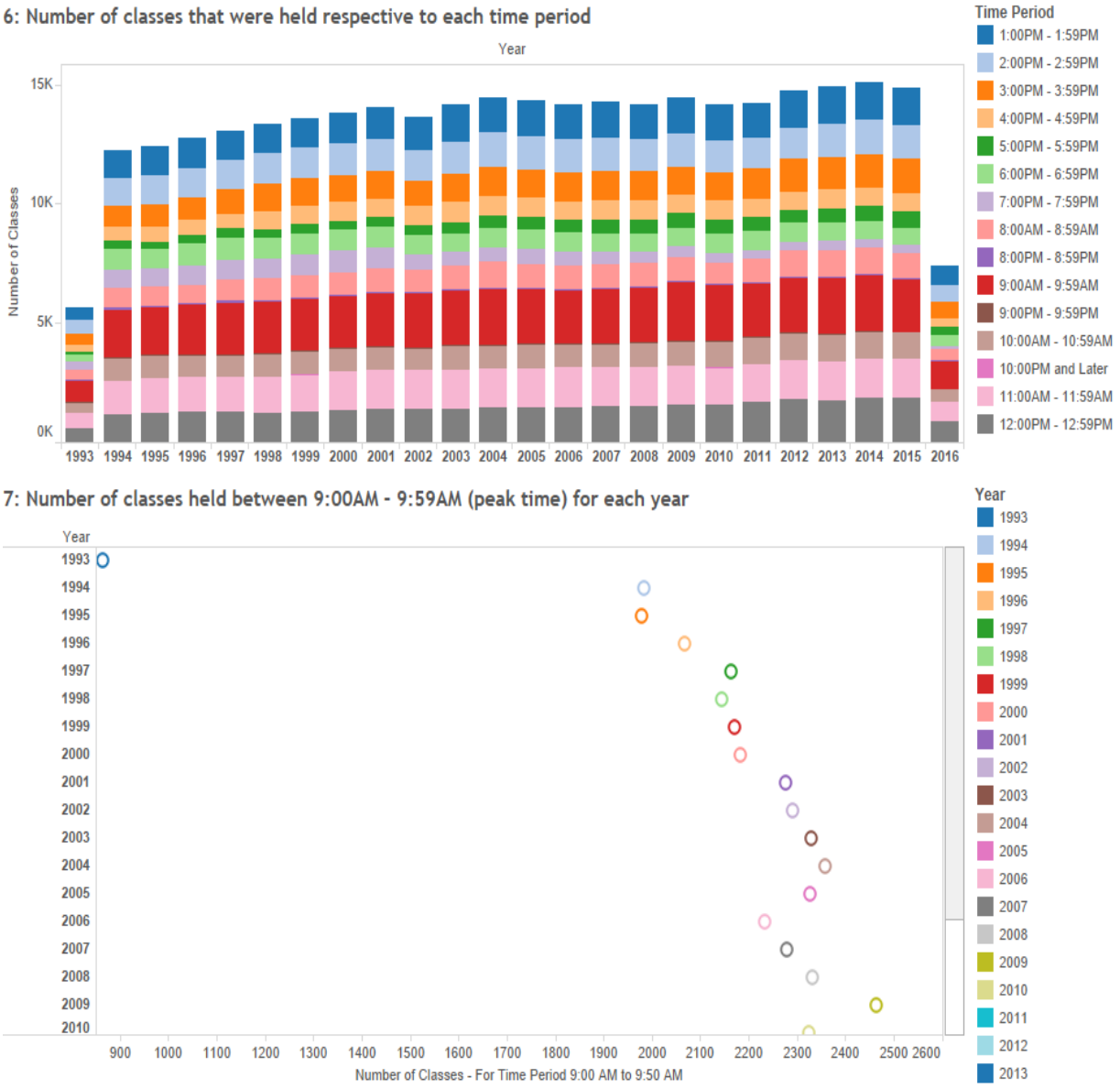


Fig 7: Scatter plot for number of classes held between 9:00AM – 9:59AM (peak time) for each year


6: Number of classes that were held respective to each time period
Year
Number of Classes
15K
10K
5K
0K
1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Time Period
1:00PM - 1:59PM
2:00PM - 2:59PM
3:00PM - 3:59PM
4:00PM - 4:59PM
5:00PM - 5:59PM
6:00PM - 6:59PM
7:00PM - 7:59PM
8:00AM - 8:59AM
8:00PM - 8:59PM
9:00AM - 9:59AM
9:00PM - 9:59PM
10:00AM - 10:59AM
10:00PM and Later
11:00AM - 11:59AM
12:00PM - 12:59PM
7: Number of classes held between 9:00AM - 9:59AM (peak time) for each year
Year
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600
Number of Classes - For Time Period 9:00 AM to 9:50 AM
Year
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

Dashboard 3

Dashboard 4: Analyzing the scheduling of all the halls based on capacity for the year 2015 and also based on efficient utilization for each semester in each year since the year 2000.

MR Code Implementation: This dashboard infers results obtained from 6 different MR Jobs of which first 3 MR Jobs constitute of first question: the top 5 halls having the maximum capacity for the recent year (Problem 6) and the next 3 MR Jobs constitute the second question: Show which halls were most efficiently utilized for each semester in each year since the year 2000 (Problem 8).

Job 1: To find the capacity for each hall.

Job 2: To sort the records on values.

Job 3: To display top 5 records.

Job 4: Finds how efficiently utilized each class was used for each semester year. Here efficiency is a measure being calculated on the basis of difference between max capacity-registered students. Lesser the difference more efficiently has the class been used.

Job 5: Finds out how many classes were there in each hall which were efficiently utilized and filters out those entries whose difference between Max capacity- registered students is less than or equal to 25.

Job 6: Finds a hall for each year with the maximum number of most efficiently utilized rooms.

From fig 8shown below, we can observe that Nsc Hall has the maximum capacity in the UB campus. Hence comparatively greater classes must be being held in this hall or this hall must be utilized for conducting events or seminars attended by huge people. Besides Baldy, knox and Dfn Hall also have greater capacity, as we all know Baldy hall has a Gym which required comparatively more space. As we have seen from Problem 3 Baldy hall has utilized the maximum time overall, also it has a comparatively greater capacity which might mostly be the reason for it being used for a greater time period. The halls with least capacity are Roosevelt and Hayes.

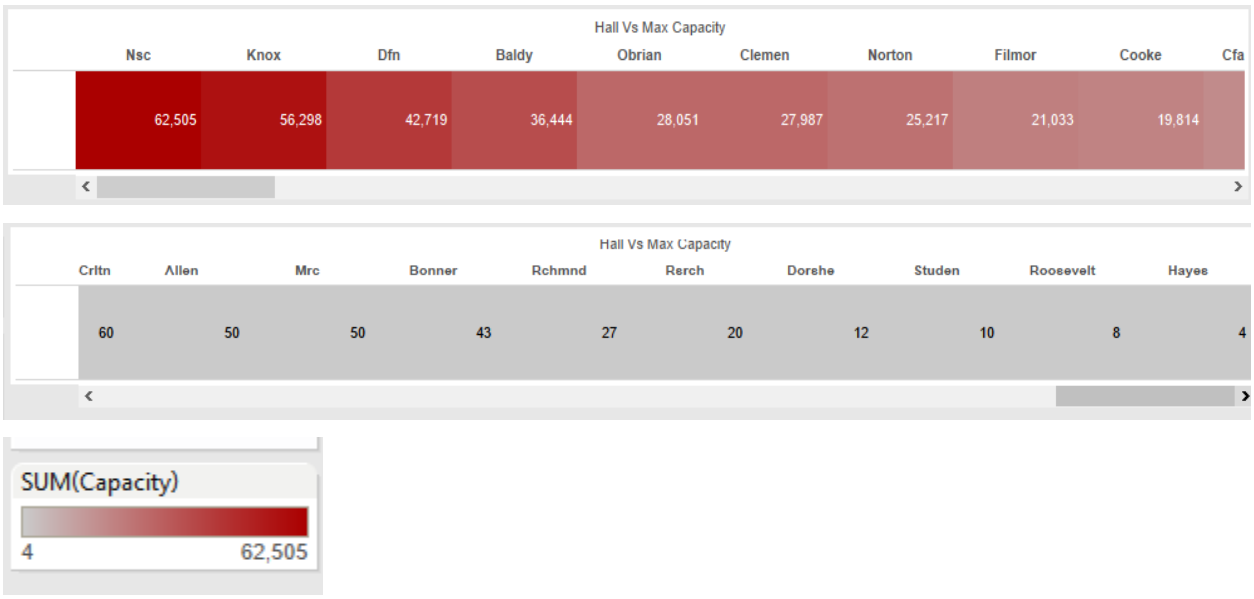


Fig 8: Highlight Table for displaying all the halls based on their capacity.

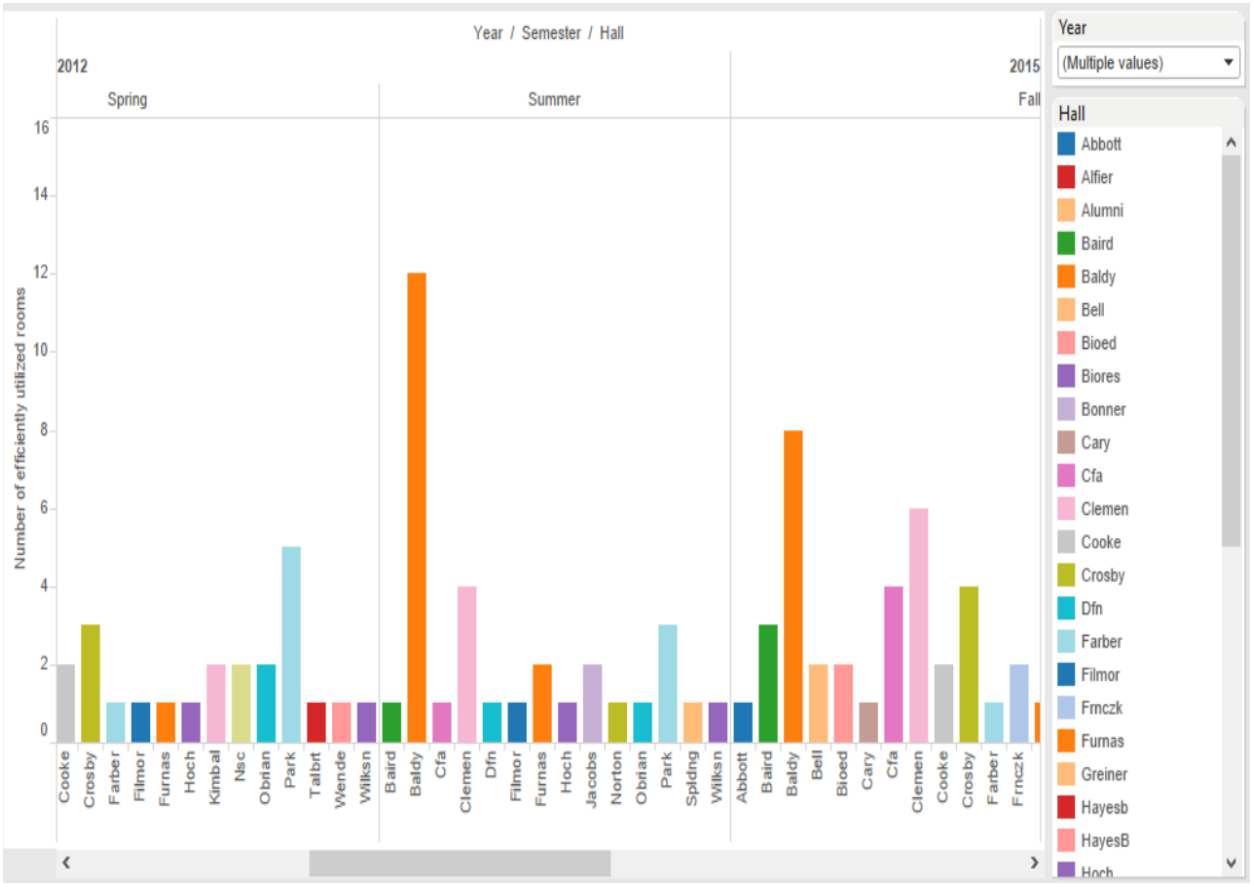


Fig 9: Bar graph showing Number of efficiently utilized rooms in each hall for each semester and year > 2000

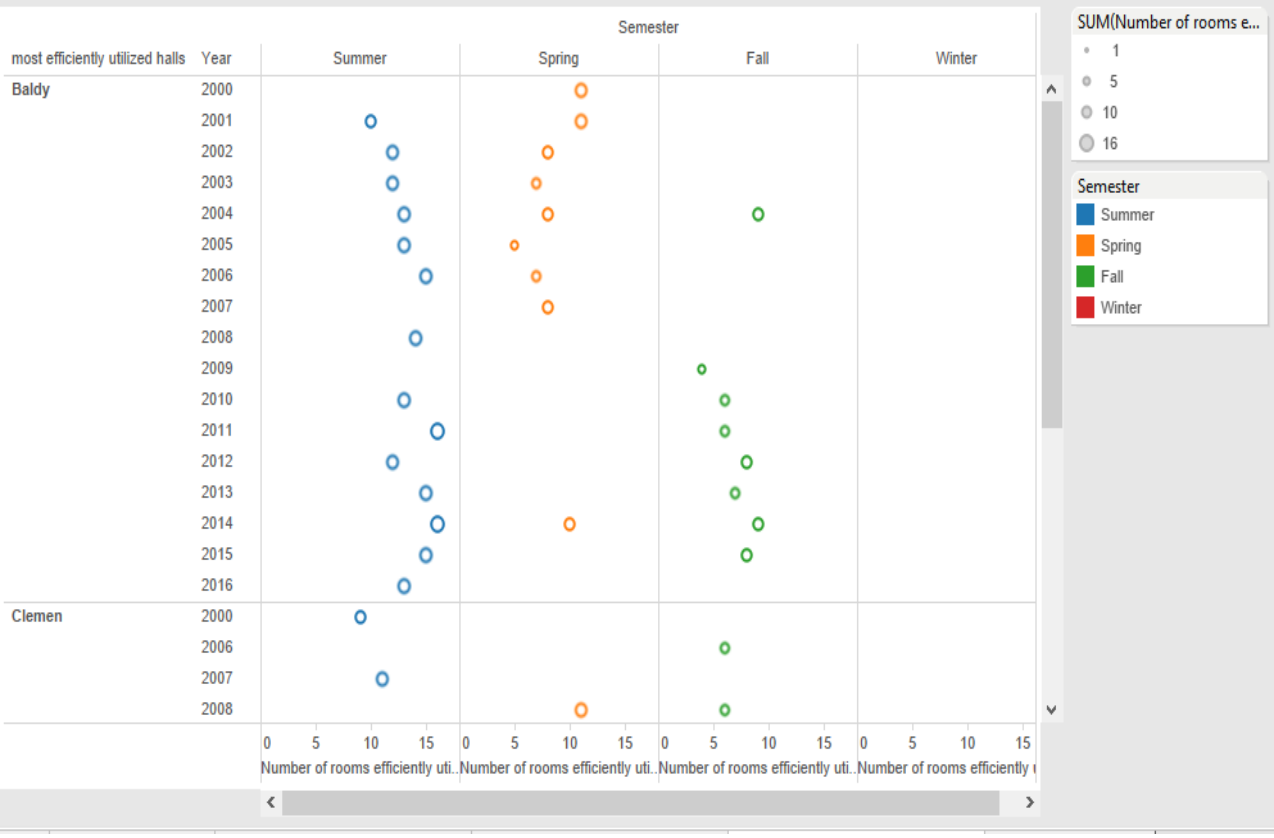
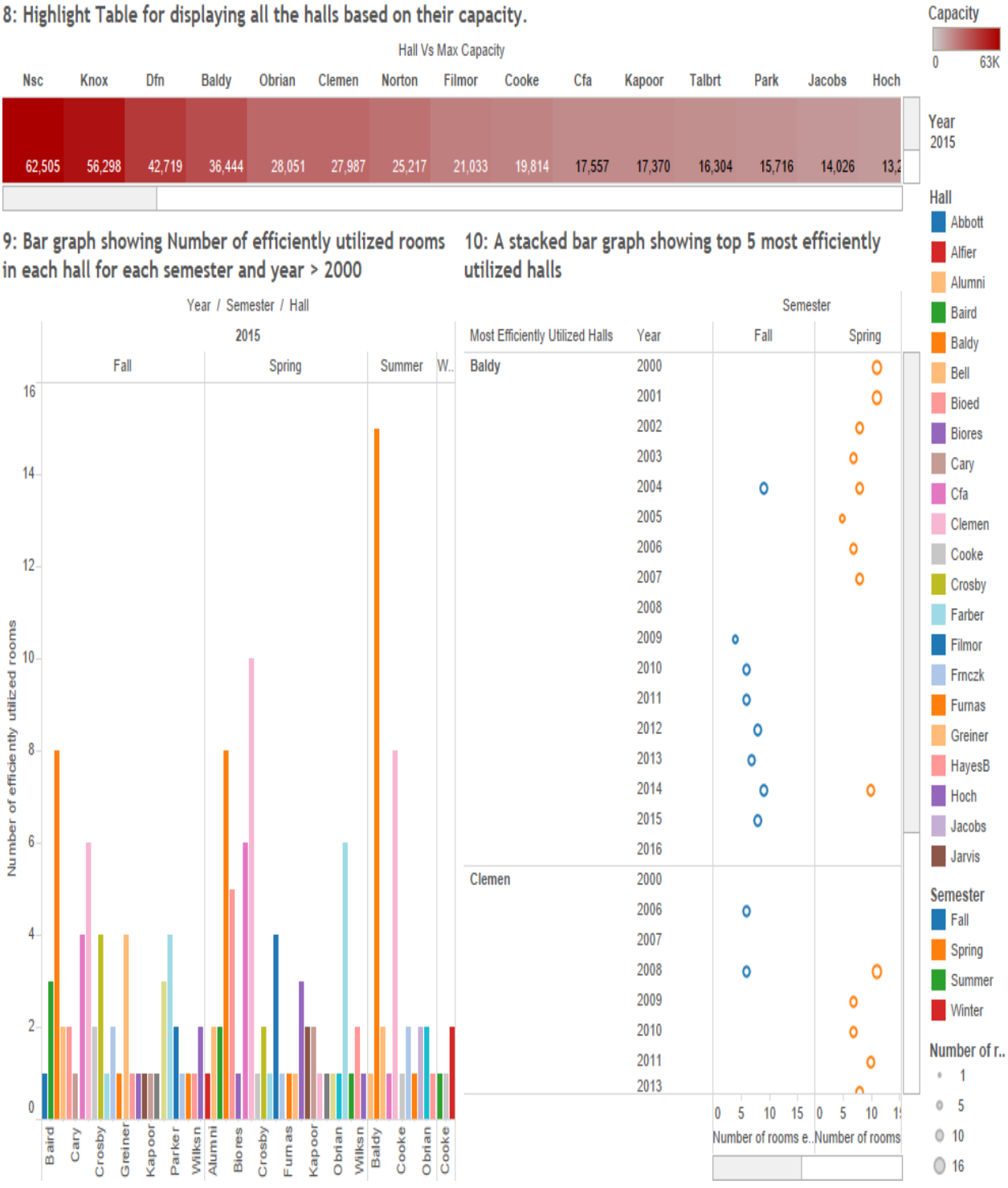


Fig 10: A stacked bar graph showing top 5 most efficiently utilized halls

On observing the visualization output we can see that the most efficiently utilized hall was Baldy over the years. This gives a clear idea about the how efficiency of the hall varied each year and in which year which hall was the most efficiently utilized. Also we can see from the above output that Baldy is most frequently most efficiently utilized hall for each semester. Also we can see that the halls are more efficiently utilized in the fall semester comparatively, and also maximum students register during this semester thus being the causes for halls being mostly efficiently utilized.


8: Highlight Table for displaying all the halls based on their capacity.
Hall Vs Max Capacity
Nsc Knox Dfn Baldy Obrian Clemen Norton Filmor Cooke Cfa Kapoor Talbrt Park Jacobs Hoch
62,505 56,298 42,719 36,444 28,051 27,987 25,217 21,033 19,814 17,557 17,370 16,304 15,716 14,026
Capacity
0 63K
Year
2015
Hall
Abbott Alfier Alumni Baird Baldy Bell Bioed Biores Cary Cfa Clemen Cooke Crosby Farber Filmor Frnczk Furnas Greiner HayesB Hoch Jacobs Jarvis
Semester
Fall Spring Summer Winter
Number of r..
1 5 10 16
9: Bar graph showing Number of efficiently utilized rooms in each hall for each semester and year > 2000
Year / Semester / Hall
2015
Fall Spring Summer W..
Number of efficiently utilized rooms
0 2 4 6 8 10 12 14 16
Baird Cary Crosby Greiner Kapoor Parker Wilksn Alumni Biores Crosby Furnas Kapoor Obrian Wilksn Baldy Cooke Obrian Cooke
10: A stacked bar graph showing top 5 most efficiently utilized halls
Semester
Most Efficiently Utilized Halls Year Fall Spring
Baldy 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Clemen 2000 2006 2007 2008 2009 2010 2011 2013
0 5 10 15 0 5 10 15
Number of rooms e.. Number of rooms

Dashboard 4

Dashboard 5: Trends and comparison of the number of students registered for online course v/s those on campus for a particular course for each year.

MR Code Implementation: This dashboard infers results obtained from 2 different MR Jobs of which basically the first find's out the number of students registered for online and on campus course (for each course respectively) and the second filters out those courses which offer both online and on campus courses (Problem 7).

Job 1: To find the number of students registered for online and on campus course (for each course respectively).

Job 2: Filtering those courses which offer both online and on campus course for that year.

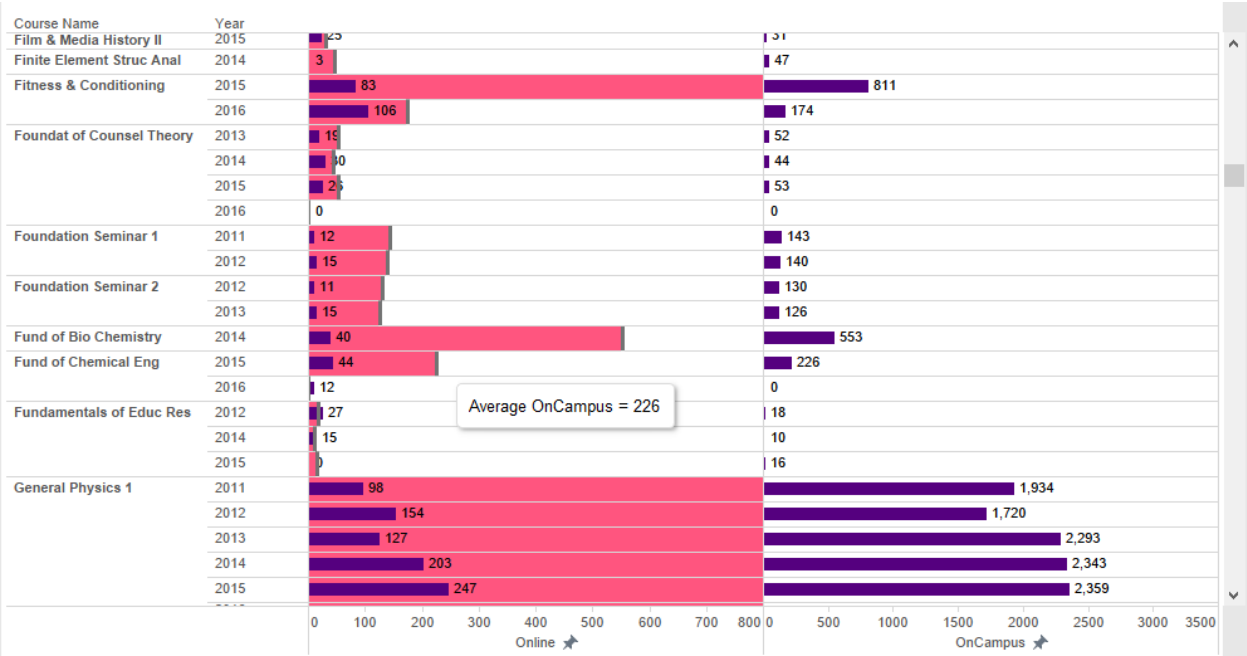


Fig 8: Bullet graphs to show comparison of students registered for online courses v/s OnCampus courses for each course respectively (purple-On line, pink- OnCampus courses in the left side, left side purple-On campus courses)

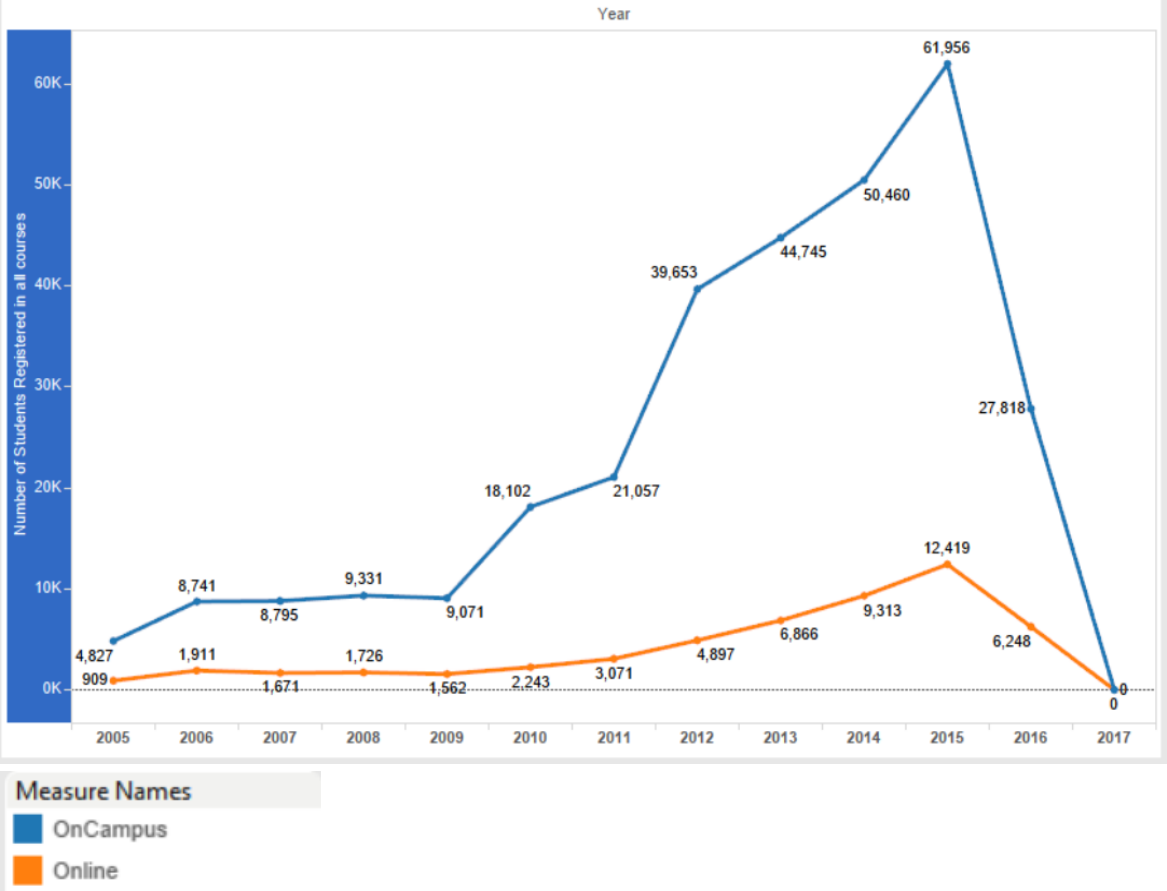
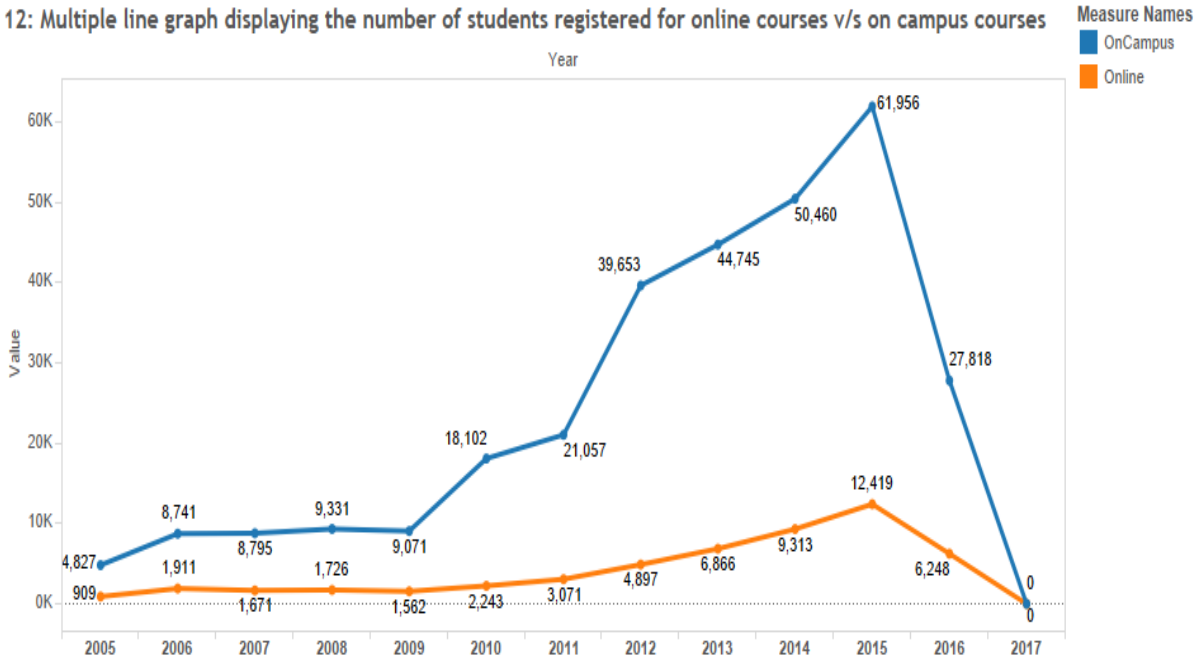


Fig 9: Multiple line graph displaying the number of students registered for online courses v/s on campus courses

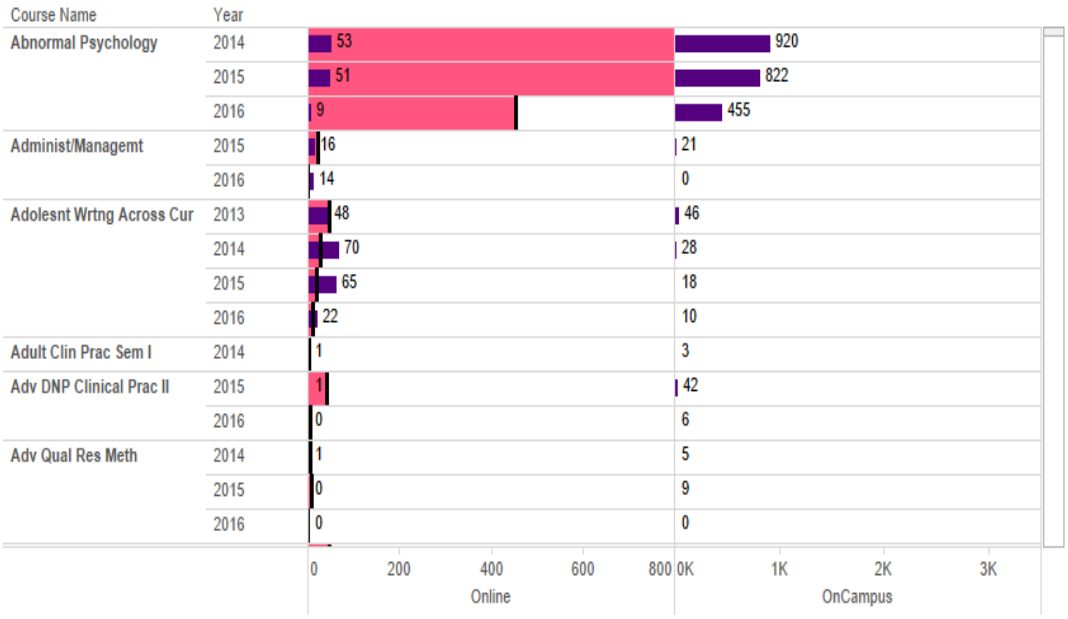
Fig 8 and Fig 9 give us a clear idea about which course had more demand online or on campus for each year. As we can see the output above for the subject Collection development in 2014 the demand for online courses is greater as compared to off campus and greater number of students are registering online. Thus allocating larger space classrooms such courses will probably not be wise decision. However, some courses like College calculus have greater demand on campus hence they should positively be considered while allocating classes. This surely will give us a clear idea about the courses whose online demand is increasing, and probably increasing the capacity would help as people are considering taking up the online course option. From fig 9 we can clearly conclude that generally there is greater demand for on campus course's than online courses. There was a sudden rise in on line courses in the year 2014-2015 and the demand had a fall since 2015! The cause for sharp decrease in the number of students registered for on line and on campus

courses in 2017, might be that the admission process for 2017 has just begin and hence the entries for 2017 are incomplete!

12: Multiple line graph displaying the number of students registered for online courses v/s on campus courses



11: Bullet graphs to show comparison of students registered for online courses v/s On campus



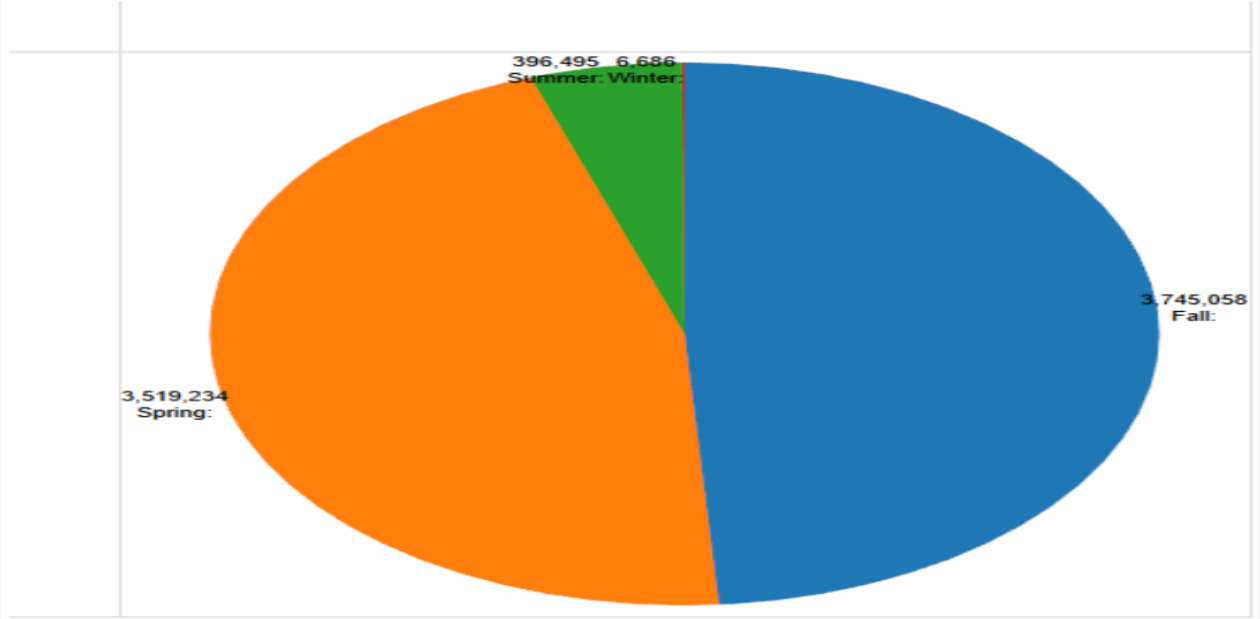
Dashboard 5

Dashboard 6: Distribution of number of Student registered over each semester.

MR Code Implementation: This dashboard infers results obtained from 2 different MR Jobs of which basically find's out the semester for which greater number of students register generally and the second one filters those courses that offer both online and on campus courses for a particular year (Problem 9).

Job 1: To get the number of students registered for Fall, Winter, Summer and Spring.

Job 2: To get the semester for which students have enrolled in greater numbers i.e. spring, fall, winter or summer etc.



Dashboard 6 & Fig 10: Pie chart Distribution for number of students registered for each semester

From the above pie chart distribution, it is clear that Greater number of students have registered for Fall semester. The Fall semester must be probably offering greater number of courses or having greater capacity. Hence many students must be registering for classes in Fall as compared to Spring, Winter and Summer. The classification for Winter semester in the pie chart is not easily visible as very less students i.e. a number of 6,686 have registered for courses in winter semester.

Story: Analysis and Visualization of Class scheduling data at UB

Following is the conclusions we can draw from the story presented in tableau in stage 3 of this project.

- General Chemistry, World Civilization 1 and World Civilization 2 are the top3 courses with maximum demand at UB in the order given i.e. General Chemistry has the most number of students registered till date. Maximum students registered (5397 students) for this course in Fall 2015.

- UB is also famous for its online courses as well, many students register for some subjects for online programs and the demand for online courses has increased over all however students do prefer on campus courses.

- Focusing on the hall utilization point of view, Baldy, Nsc and Clemen are the top 3 halls that utilized most amount of time as compared to other halls and Hall Hadly was utilized for very less time comparatively i.e. only one hour. Baldy Hall which was most utilized time wise, saw a sudden rise in the number of hours utilized in the year 2016-2017 of about 1,762 hours. Nsc, Knox, Baldy, Dfn and Obrian are halls having greater capacity as compared to others out which Baldy is most efficiently utilized in almost each year.

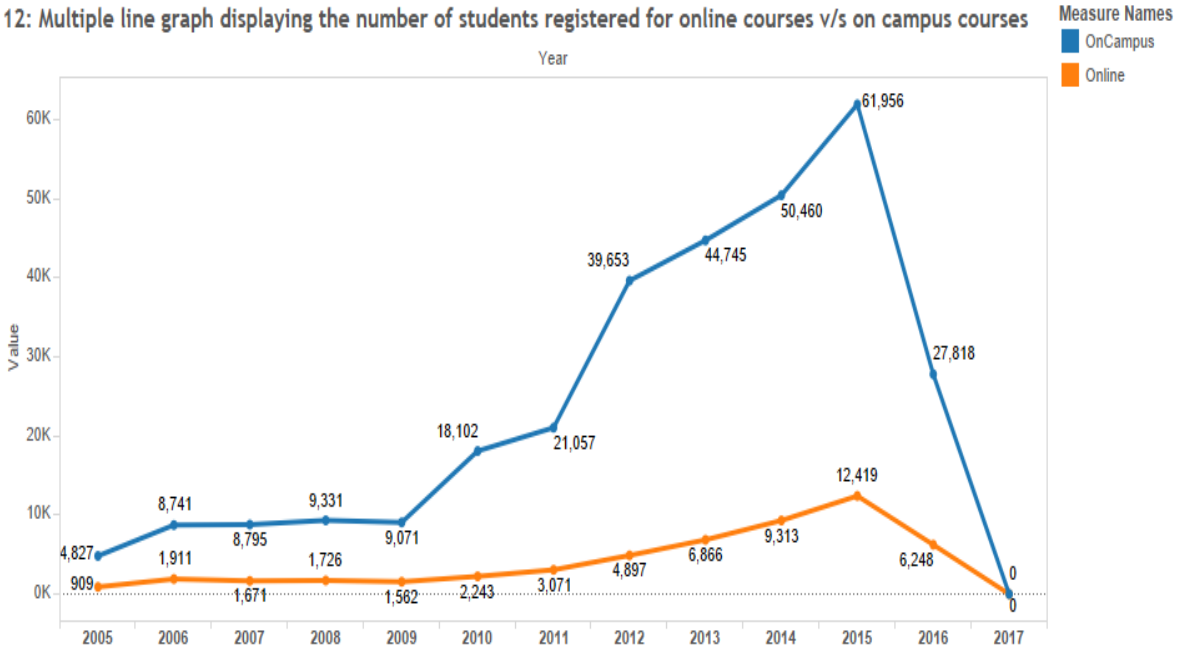
- The peak time for which most classes were held each year was 9:00 AM to 9:50 AM and hence the classrooms in UB campus seems to be occupied at morning time specially at 9 a' clock in the morning. Maximum number of classes were held during this period in the year 2009.

- Fall semesters seem to be preferred by students as these semesters experience maximum number of student registration.

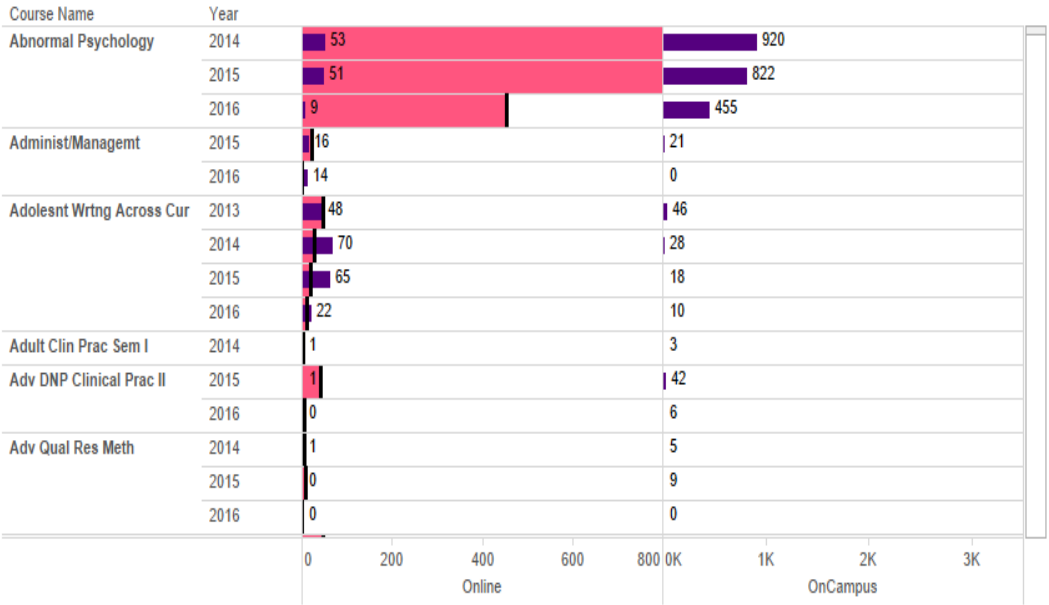
Analysis & Visvualisation of Class Scheduling data at UB

Top 20 courses for which maximum students have registe..	Hall that utilized maximum time since year 2000 and trends..	Peak time for which most number of classes were held for..	Analyzing the scheduling of all the halls based on capac..	Trends and comparison of the number of students r..	Distribution of number of Student registered over each semester

12: Multiple line graph displaying the number of students registered for online courses v/s on campus courses



11: Bullet graphs to show comparison of students registered for online courses v/s On campus



Story: Analysis and Visualization of Class scheduling data at UB (Dashboard 5)