CSE587

Data-intensive Computing

Spring 2016

Team members

Name : Harsha Sudarshan
UB Person Number: 50169593

Name : Jayneesh H Wanjara
UB Person Number: 50169911

Problem 2: Design of questions (Experiments) Report

Introduction:

This report consists of the various questions that were formulated and implemented in order to provide insights into the situation of classroom scheduling at the UB Campus. We have chosen to answer 5 complex questions which include multiple MR stages, since we felt that we would be able to do more justice to the data and its various dimensions.

We have elaborated below the reasoning behind the questions and the implementation strategy that best provides valuable insights into the classroom allocation and scheduling problem. The sections will also include the contribution of the question to the overall problem, the overall stages of the MR implementation and the format of the output. We have also specified which file was used as the input.

The data cleaning process eliminated every cell in the CSV file that included values such as “Arr”, “arr”, “UNKWN”, “unknown” in its fields. Also classes that say “Online” on the location field have not been considered since they probably did not need a classroom.

Question 1:

The first question that we have designed could be framed as follows -
“For every Semester, provide the most and least used room, for every working day of the week (Monday to Saturday/Friday). Also for each day of the week, across all the semesters, provide the average number of classes held on that day.”

The output of this question will be in the following format(“Key” “Value”)
“Semester ~ day of the week” “least Number of classes conducted on this day ~ corresponding hall name and room number ~ Highest number of classes conducted on this day of the week ~ corresponding hall name and number ~ average number of classes conducted throughout the semester on this day”

Sample Output Line:

```
Fall 1993~F      0001~Kimbal 1140~0008~Talbrt 10~0004
```

Fields of the output :
Fall 1993→ The semester
F → Friday. The day of the week (M,T,W,R,F,S,U)
0001 → The count of the class
Kimbal 1140→ The name of the room with the least number of classes (HallName

RoomNumber)
0008 → The count of the class
Talbrt 10 → The name of the classroom with the most number of class for Friday
0004 → The average number of classes conducted on Friday for the semester Fall 1993

The output sample given above basically tells us that the room Kimbal 1140 was used once every Friday for the semester fall 1993, and the room Talbrt 10 was used 8 times every Friday in the same semester. And the average classes held every Friday of that semester was 4.

Special Note
Some classes in the semester have 2 entries for the same time slot in a day. These instances have been counted only once, since we cannot have 2 classes held in the same room in the same timeslot.

Example:

Fall 1993	Talbrt 10	M	7:00PM - 7:59PM	6382	Intro to Philosophy	29	60
Fall 1993	Talbrt 10	M	7:00PM - 7:59PM	6382	Intro to Philosophy	29	60

As you can see in the above example taken from the input file, there are two entries for the same class, held on the same day, at the same time in the entire semester. There are instances like this quite often in the input file, they have been considered only once.

Input file:
For this question, we have considered the mail CSV file - bina_classschedule.csv as the input file.

Question 2:

The second question that we have framed can be framed as follows:
"Provide the busiest and the leanest timeslot of the day, for all the four semester, across all the documented active years. Also provide the overall average of the number of classes for the all the time slots of the day, across all the documented years."

The output of this question will be in the following format("Key" "Value"):
"Semester" "Count of the number of classes ~ corresponding time slot of the day that was the least used ~ count of the classes ~ Corresponding time slot of the day that was busiest~ average number of classes for all timeslots across the years"

Sample Output Line :

Fall 0002~10:00PM and Later~8573~4:00PM - 4:59PM~9394
Fall → The name of the semester (across all the years)
0002 → The number of classes held
10:00PM and Later → The most idle timeslot of the day for that semester (fall)
8573 → Number of classes held
4:00PM - 4:59PM → The busiest timeslot of the day for the fall semester
9394 → The average number of classes held over all the timeslots on an average for the semester

The output basically says that over the years , the fall semester was the most idle(had the least number of classes) during the timeslot , 10:00PM and Later and it was the busiest during the timeslot, 4:00PM - 4:59PM. Also across all the fall semesters, the average number of classes held was 9394.

Special Note

In this particular question, segregated classes held on more than one day of the week, as different instances. For example,

Fall 1993	Talbrt 10	MW	8:00AM - 8:59AM	4061	Human Resources Mgt	50	60

In the above entry of the input file, we can see that this course was held on Monday and Wednesday, from 8:00AM – 8:59AM, so we have counted these as two instances when we performed the count, since they are busy twice during that week.

Input file

The input file for this question was the original CSV file, bina_classschedule.csv

Question 3:

The third question that we have designed could be framed as follows -

‘For each Hall (inclusive of all the rooms) provide the percentage utilization of that room, according to its maximum usage capability. Then list the rooms with the highest and lowest utilization for that Hall. Also provide the average utilization percentage including all the rooms for that Hall”

How the percentage is calculated:
Across all the years and semesters , the maximum enrolled class size was found, this was assumed to be the maximum limit of the class (Since the maximum capacity was not found to be accurate and every course barely uses the exact class size completely). Once the maximum class size for a room was found, we checked the various class sizes that the same room accommodated. This number was found for every semester, and then was divided with the maximum capacity of that room. This was the percentage utilization of that room. This was done for all the halls, and for every hall, we displayed the classes with the least and maximum utilization percentage. Also the average class utilization was calculated for every Hall.

The output of this question will be in the following format (“Key” “Value”):
“Name of the hall” “Percentage Utilization_Room least utilized_percentage utilization_room best utilized_average percent utilization of all the rooms in this hall”.

Sample Output line:
Abbott 0033_165_0065_B-15_0044

Fields of the output:
Abbott → Name of the Hall
0033 → The percentage utilization of the corresponding room
165 → The room number that was least utilized in Abbott hall
0065 → The percentage utilization of the corresponding room number
B-15 → The room number with the maximum utilization in Abbott hall
0044 → the average percentage utilization of all the rooms in Abbott Hall

The output basically says that for Abbott hall the room Abbott 165 is the least efficiently used over the years and the room Abbott B-15 was the most efficiently used from the beginning up to this point.
The average utilization of the rooms for Abbott hall is quite low at 44%.

Special Note:
In this particular scenario we have excluded those classes where the strength of the student was zero, since these are cancelled classes and probably never ended up using the classroom anyway.

Input file:
The input file for this question is bina_classschedule.csv

Question 4:

The fourth question that we have designed could be framed as follows:
“Across all the years , for every department of UB, provide the count of percentile of the classes for different range of the classroom sizes. “

The format of the output file is as follows (“Key” “Value”):
“Abbreviation of the department~Percentile”   “Range of the classroom size”

Sample Output Line:

```
AAS~002         060-099
```

Fields of the output line:
AAS → Abbreviation of the department name (African and African American Studies)
002 → Percentile of the classes
060-099 → range of the classrooms based on the number of student enrolled for each course

This output example, says that 2 percent of the overall classes of AAS have a class size of 060 to 099 students that have enrolled for these classes.

Input file:
For this particular question, we have used bina_classschedule2.csv as the input file.

Special Note:
For any given department the sum of all the percentiles for the various ranges will be approximately 100.

Question 5:

The fifth question can be framed as follows:
“Provide the percentile for the size for the enrolled class size for every year. “

The format of the output file is as follows (“Key” “Value”):
“Academic Year~Batch Size of students”   “Number of classes that year which fell under this range”

Sample output Line:

```
1993~000-020	046
```

Fields of the output line:
1993 →Academic year
000-020 → Range of the number of students enrolled for any given class offered that year
046 → Percentile of students that have enrolled that year , for any class that falls within that range

The above line basically say that for the year 1993, 46 percent of the total students that enrolled in UB were part of a class that falls in the size of 000-020.

Input file:
The input for this question was the main CSV file, that is, bina_classschedule.csv

Conclusion:

As you can see we have framed the questions in such a way that all the answers will help us collectively to solve the problem of class utilization by giving a better, clearer picture of the demand and allocation.
Also as you will see in the code itself, we have saved the output of the intermediate stages separately, so as to use them in the next phase of this project on Tableau.

With the help of this intermediate data and the above collected output files, we will provide some insightful graphs that can enable the reader to help understand and visualize the issue of student accommodation into the classrooms that best fit the criterion.