

# Database warehousing, Integration and Provenance-CS520

## Vizier assignment

### Group 10

Shivaji Goud Panam

spanam@hawk.iit.edu

Gowtham Kumar Kamuni

gkamuni@hawk.it.edu

Sirisha Gandham

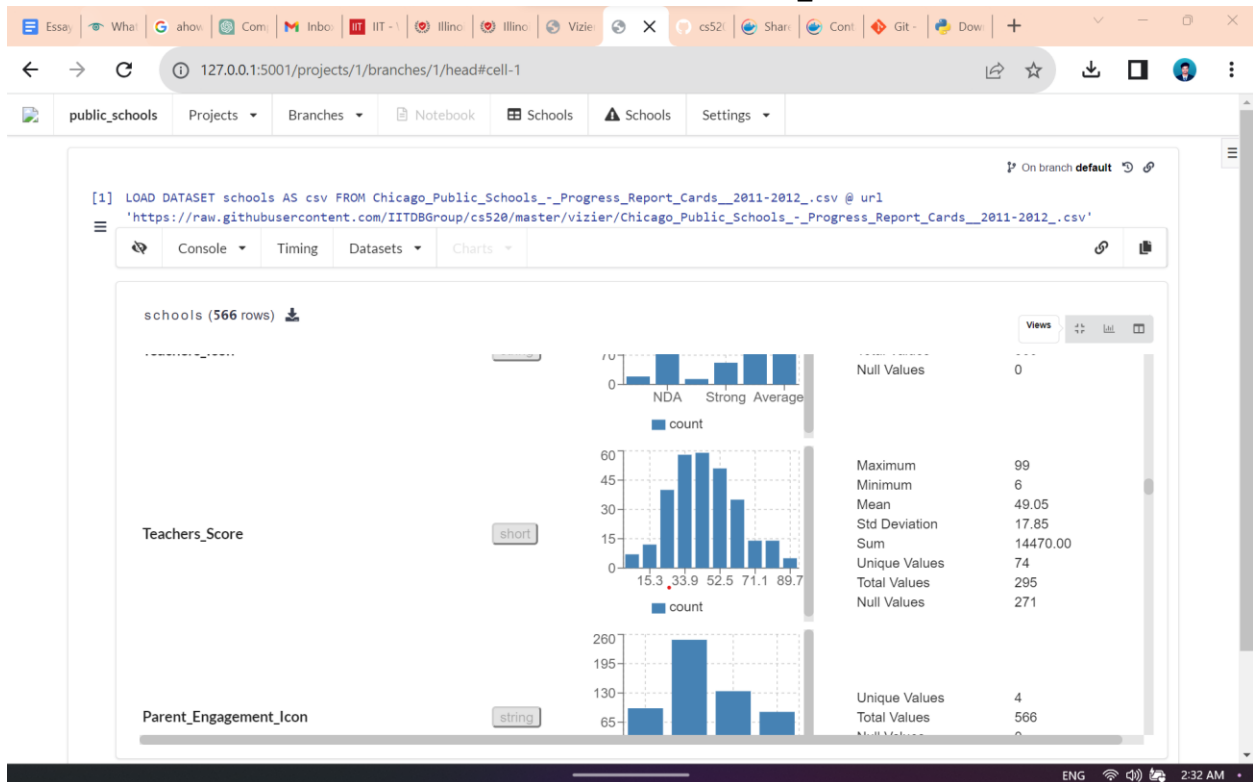
sgandham@hawk.it.edu

**Task 1:** Load a dataset and take a screenshot of the result

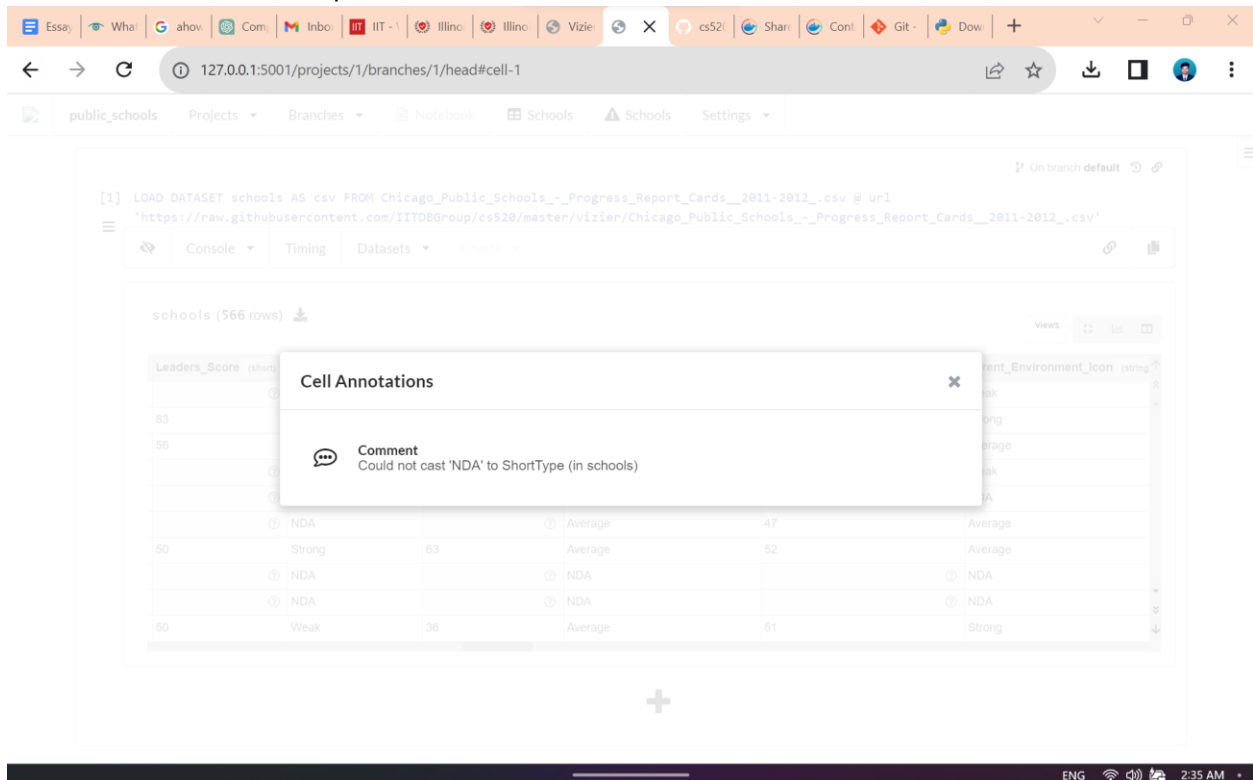
The screenshot shows a web browser window with a URL bar displaying '127.0.0.1:5001/projects/1/branches/1/head'. The browser has several tabs open, including 'Essa', 'Whe', 'aho', 'Con', 'Inbc', 'IIT', 'Illin', 'Vizie', 'cs5', 'Sha', 'Con', 'Git', 'Dov', 'Gmail', and a '+' icon for more tabs. The browser's address bar shows the URL '127.0.0.1:5001/projects/1/branches/1/head'. Below the address bar, there is a navigation bar with links for 'public\_schools', 'Projects', 'Branches', 'Notebook', 'Schools', 'Settings', and a dropdown menu. The main content area displays a SQL query execution result. The query is: [1] LOAD DATASET schools AS csv FROM Chicago\_Public\_Schools\_-\_Progress\_Report\_Cards\_\_2011-2012\_.csv @ url 'https://raw.githubusercontent.com/IITDBGroup/cs520/master/vizier/Chicago\_Public\_Schools\_-\_Progress\_Report\_Cards\_\_2011-2012\_.csv'. The result is shown in a table with 566 rows. The table has columns: School\_ID (int), Name\_of\_School (string), Elementary\_Middle\_or\_High\_School (string), Street\_Address (string), City (string), State (string), and ZIP\_Coc. The table is sorted by School\_ID in ascending order. The first 15 rows are shown, with the last row (14) being the 566th row. The table is displayed in a 'Table' view. The bottom status bar shows 'ENG', a signal strength icon, a battery icon, and the time '2:27 AM'.

School_ID (int)	Name_of_School (string)	Elementary_Middle_or_High_School (string)	Street_Address (string)	City (string)	State (string)	ZIP_Coc
0 609966	Charles G Hammond Elementary School	ES	2819 W 21st Pl	Chicago	IL	60623
1 610539	Marvin Camras Elementary School	ES	3000 N Mango Ave	Chicago	IL	60634
2 609852	Eliza Chappell Elementary School	ES	2135 W Foster Ave	Chicago	IL	60625
3 609835	Daniel R Cameron Elementary School	ES	1234 N Monticello Ave	Chicago	IL	60651
4 610521	Sir Miles Davis Magnet Elementary Academy	ES	6730 S Paulina St	Chicago	IL	60636
5 609818	Luther Burbank Elementary School	ES	2035 N Mobile Ave	Chicago	IL	60639
6 610298	Lenart Elementary Regional Gifted Center	ES	8101 S LaSalle St	Chicago	IL	60620
7 610200	James N Thorp Elementary School	ES	8914 S Buffalo Ave	Chicago	IL	60617
8 609680	Walter Payton College Preparatory High School	HS	1034 N Wells St	Chicago	IL	60610
9 610056	Roswell B Mason Elementary School	ES	4217 W 18th St	Chicago	IL	60623
10 609848	Ira F Aldridge Elementary School	ES	630 E 131st St	Chicago	IL	60827
11 610038	Abraham Lincoln Elementary School	ES	615 W Kemper Pl	Chicago	IL	60614
12 610123	William Penn Elementary School	ES	1616 S Avers Ave	Chicago	IL	60623
13 609863	Christopher Columbus Elementary School	ES	1003 N Leavitt St	Chicago	IL	60622
14 610226	Socorro Sandoval Elementary School	ES	5534 S Saint Louis Ave	Chicago	IL	60629

**Task 2:** Select the detail view and look at the distributions of some columns. Then look at the column view and take a screenshot of the distribution for column Teachers\_Score.



**Task 3:** Click on one of the question marks for values in the teachers column and take a screenshot.

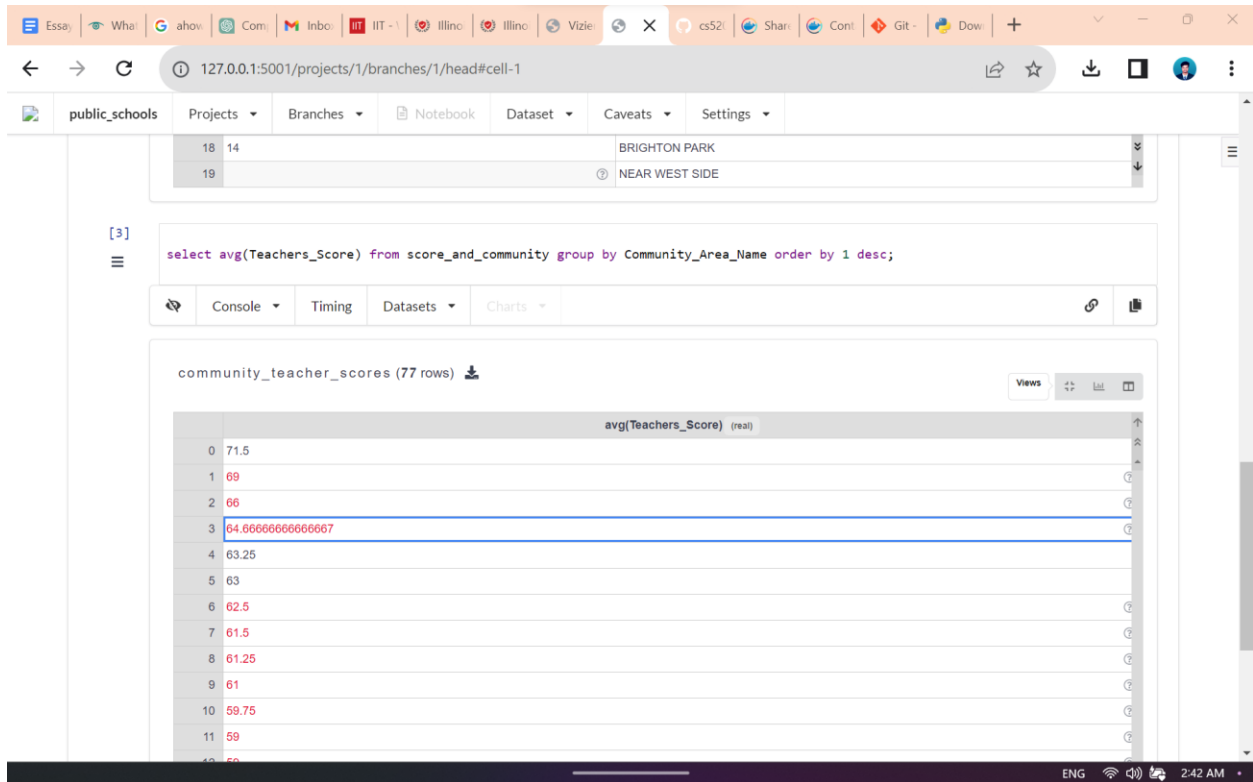


**Task 4:** Create a SQL cell and write a query that returns columns Teachers\_Score and Community\_Area\_Name. SQL results can be stored as new datasets in Vizier. Call the result dataset score\_and\_community. And take a screenshot of the result.

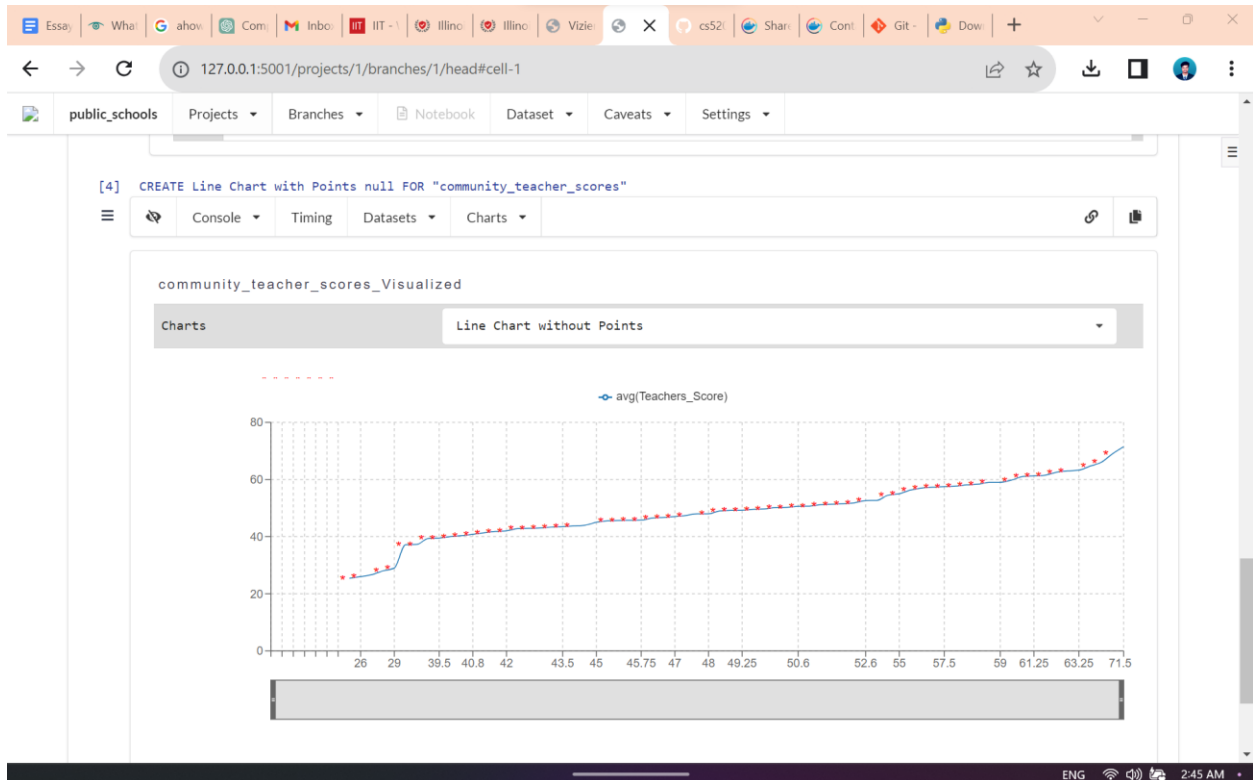
The screenshot shows the Vizier web interface. At the top, there's a navigation bar with tabs for 'public\_schools', 'Projects', 'Branches', 'Notebook', 'Dataset', 'Caveats', and 'Settings'. Below this, a SQL cell is active, displaying the query: `select Teachers_Score, Community_Area_Name from schools;`. Below the query, there's a toolbar with 'Console', 'Timing', 'Datasets', and 'Charts'. The 'Datasets' tab is selected, showing a dataset named 'score\_and\_community' with 566 rows. The dataset is displayed as a table with two columns: 'Teachers\_Score (short)' and 'Community\_Area\_Name (string)'. The table shows 13 rows of data, with the first row having an index of 0 and the last row having an index of 12. The bottom status bar shows 'ENG', signal icons, and the time '2:38 AM'.

	Teachers_Score (short)	Community_Area_Name (string)
0		SOUTH LAWNSDALE
1	88	BELMONT CRAGIN
2	48	LINCOLN SQUARE
3		HUMBOLDT PARK
4		WEST ENGLEWOOD
5		BELMONT CRAGIN
6	63	CHATHAM
7		SOUTH CHICAGO
8		NEAR NORTH SIDE
9	36	NORTH LAWNSDALE
10		RIVERDALE
11	70	LINCOLN PARK
12		NORTH LAWNSDALE

**Task 5:** Create a SQL cell and write a query over the over the score\_and\_community dataset that computes the result as described above. Call the result dataset community\_teacher\_scores. And take a screenshot of the result.



**Task 6:** Create a line chart of the aggregation result by creating a plot cell and take a screenshot of the result.



**Task 7:** Insert a new cell above the SQL cell that computes the average teacher scores (notebooks in Vizier are executed top down) by pressing the three bars below the cell number. Select "Impute Missing Values", select the score\_and\_community dataset and Teachers\_Score as the column to be imputed, and select mean as the imputation method and take a screenshot of the updated line chart.

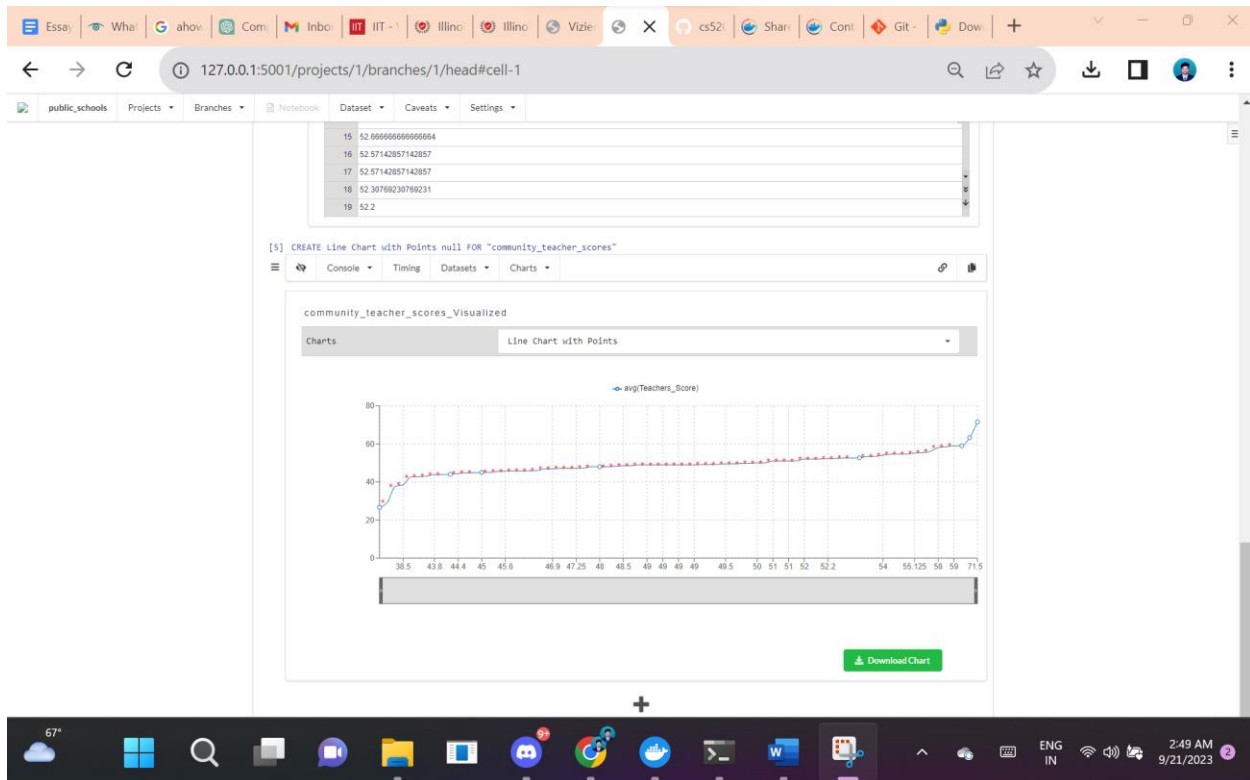
The screenshot shows the Vizier interface with a notebook titled 'public\_schools'. The notebook has two cells:

[3] `CREATE LENS ON score_and_community IMPUTE MISSING VALUES ON COLUMN @`

[4] `select avg(Teachers_Score) from score_and_community group by Community_Area_Name order by 1 desc;`

Below the cells, a table titled 'community\_teacher\_scores (77 rows)' is displayed. The table has a single column 'avg(Teachers\_Score)' and 17 rows of data. The values are as follows:

	avg(Teachers_Score)
0	71.5
1	63.25
2	59
3	59
4	58.4
5	58
6	56
7	55.30769230769231
8	55.125
9	54.666666666666664
10	54.666666666666664
11	54.5
12	54
13	53.5
14	53.333333333333336
15	52.666666666666664
16	52.57142857142857



**Task 8:** Create a Python cell at the end of the notebook and create a function called `print_avg_teachers` that uses Vizier's API to get a handle for this dataset and print all values of the `avg_teacher_score` column. Hint: use the "Show Code Examples" button to see example Vizier API usage and see here for the API documentation. Then use `vizierdb.export_module` to export the function. Then create a second Python cell and use `vizierdb.get_module("print_avg_teachers")` for importing the function and then call it. Take a screenshot of the result.

**Task 9:** Create another Python cell and use Vizier's API to access the dataset `community_teacher_scores` as a DataFrame, then filter out rows where the `avg_teacher_score` is larger than or equal to 30.0 and then print the remaining rows and take a screenshot.