Database warehousing, Integration and Provenance-CS520

Vizier assignment

Group 10

Shivaji Goud Panam

spanam@hawk.iit.edu

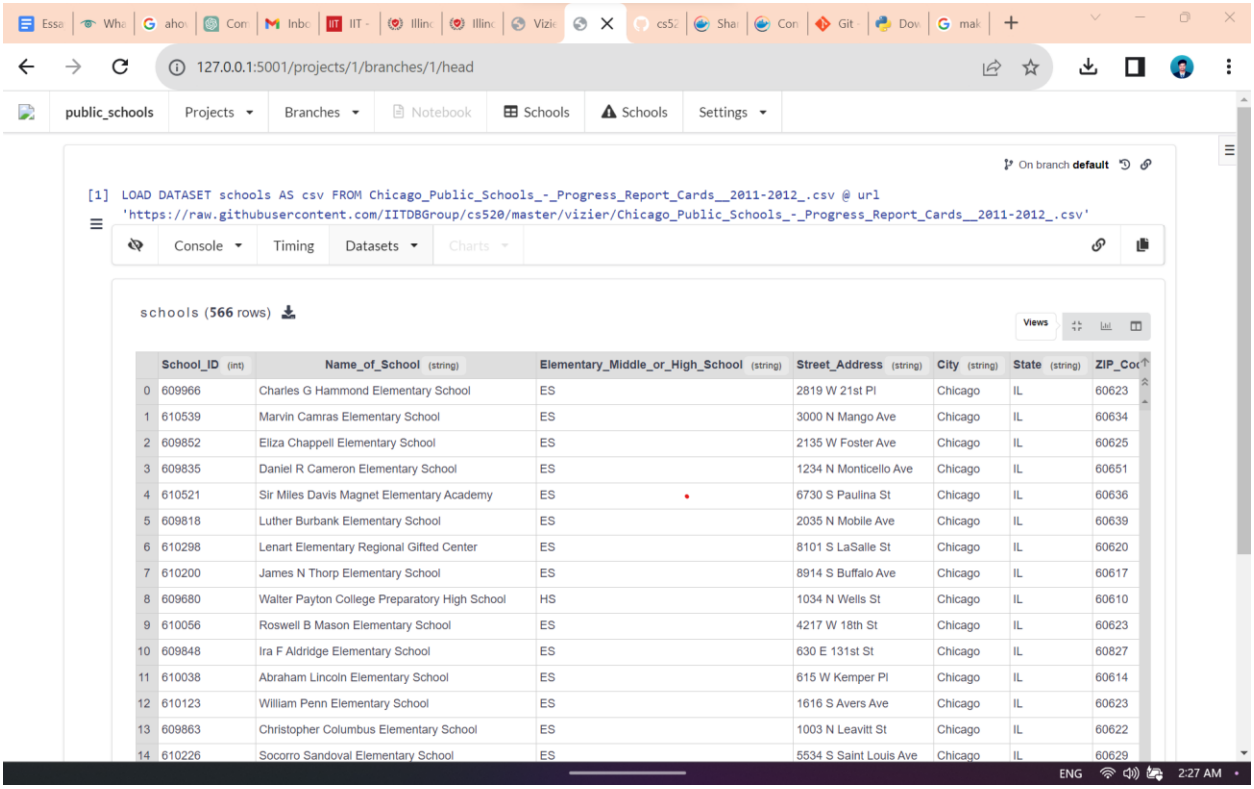
Gowtham Kumar Kamuni

gkamuni@hawk.it.edu

Sirisha Gandham

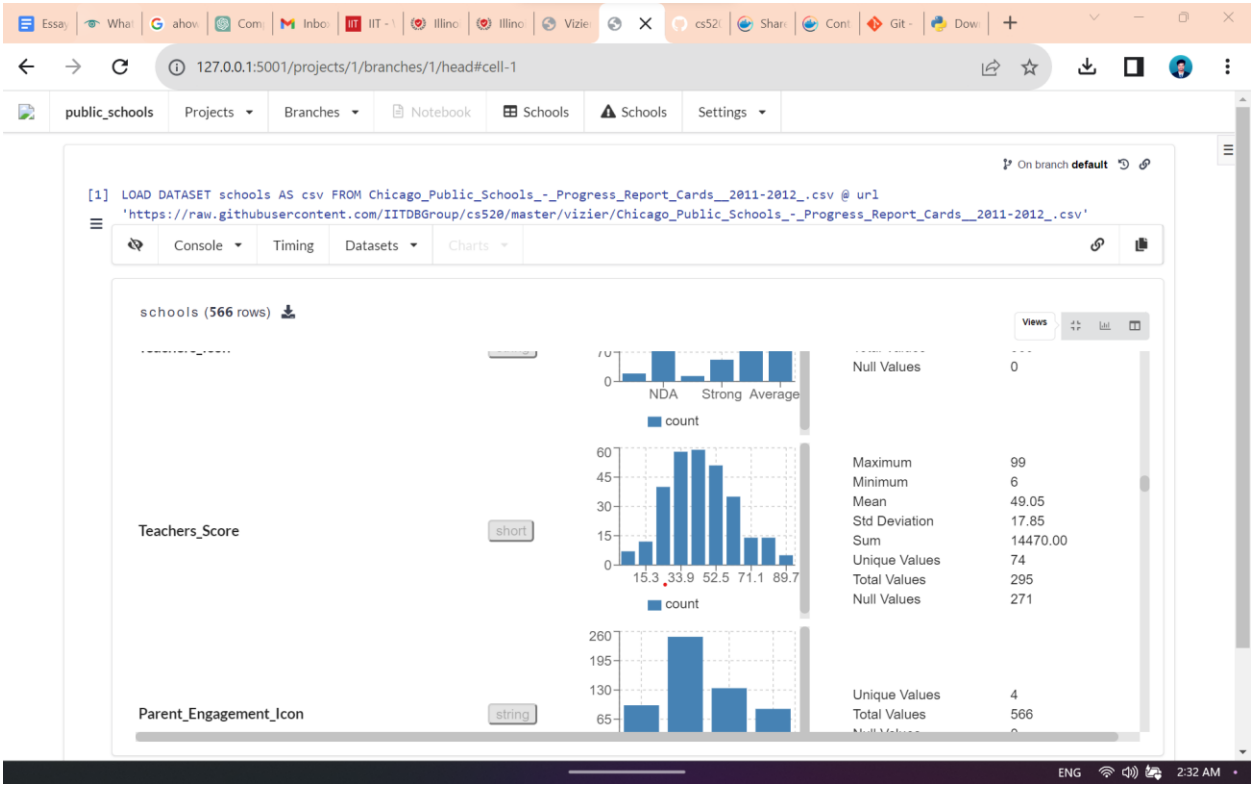
sgandham@hawk.it.edu

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

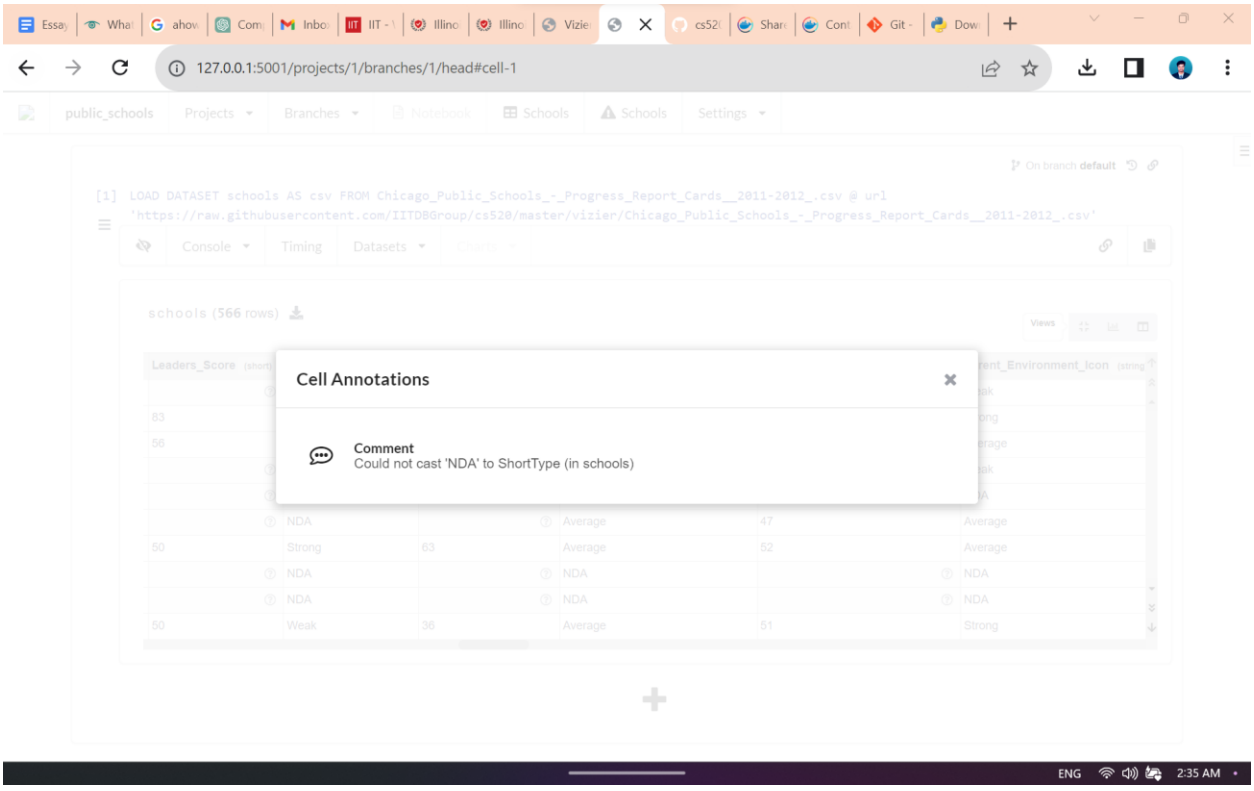


	School_ID (int)	Name_of_School (string)	Elementary_Middle_or_High_School (string)	Street_Address (string)	City (string)	State (string)	ZIP_Co
0	609966	Charles G Hammond Elementary School	ES	2819 W 21st Pl	Chicago	IL	60623
1	610539	Marvin Camras Elementary School	ES	3000 N Mango Ave	Chicago	IL	60634
2	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

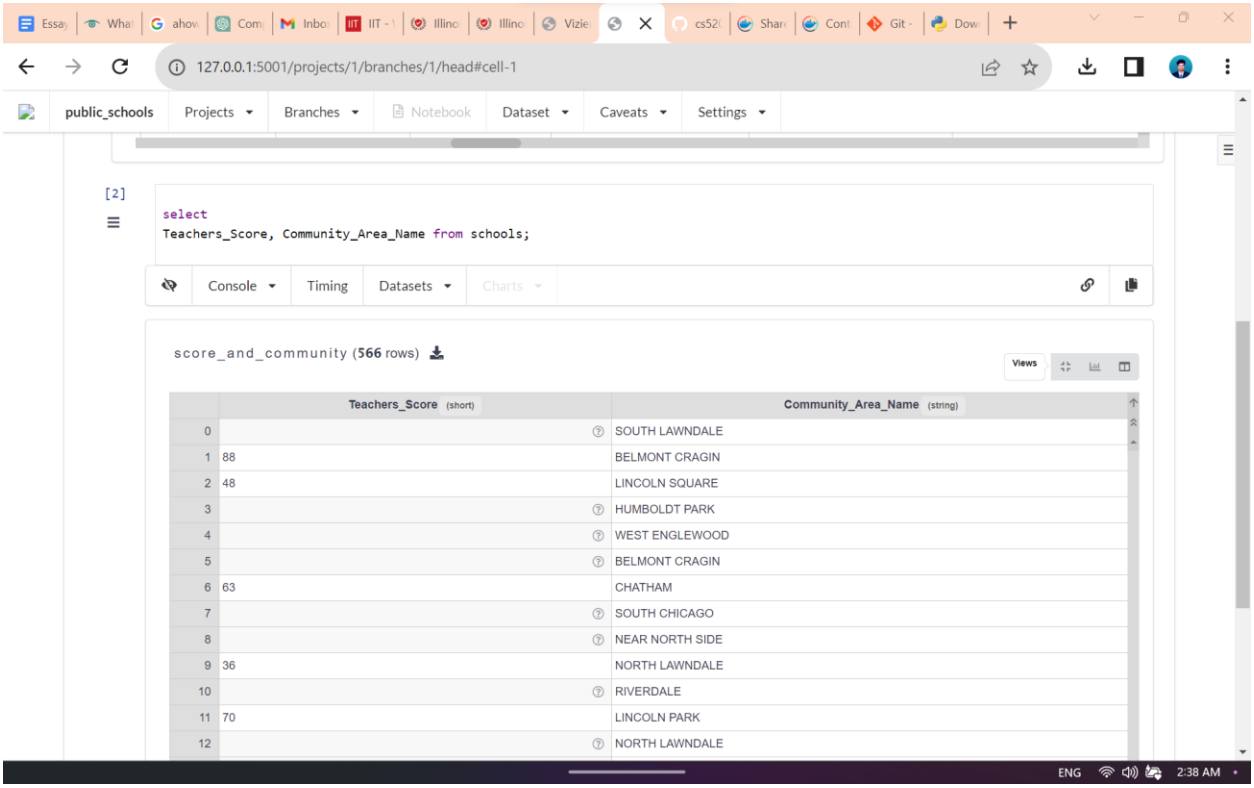
Task 2: Select the detail view and look at the distributions of some columns. Then look at the column view and take a screenshot of the distribution for column Teachers_Score.



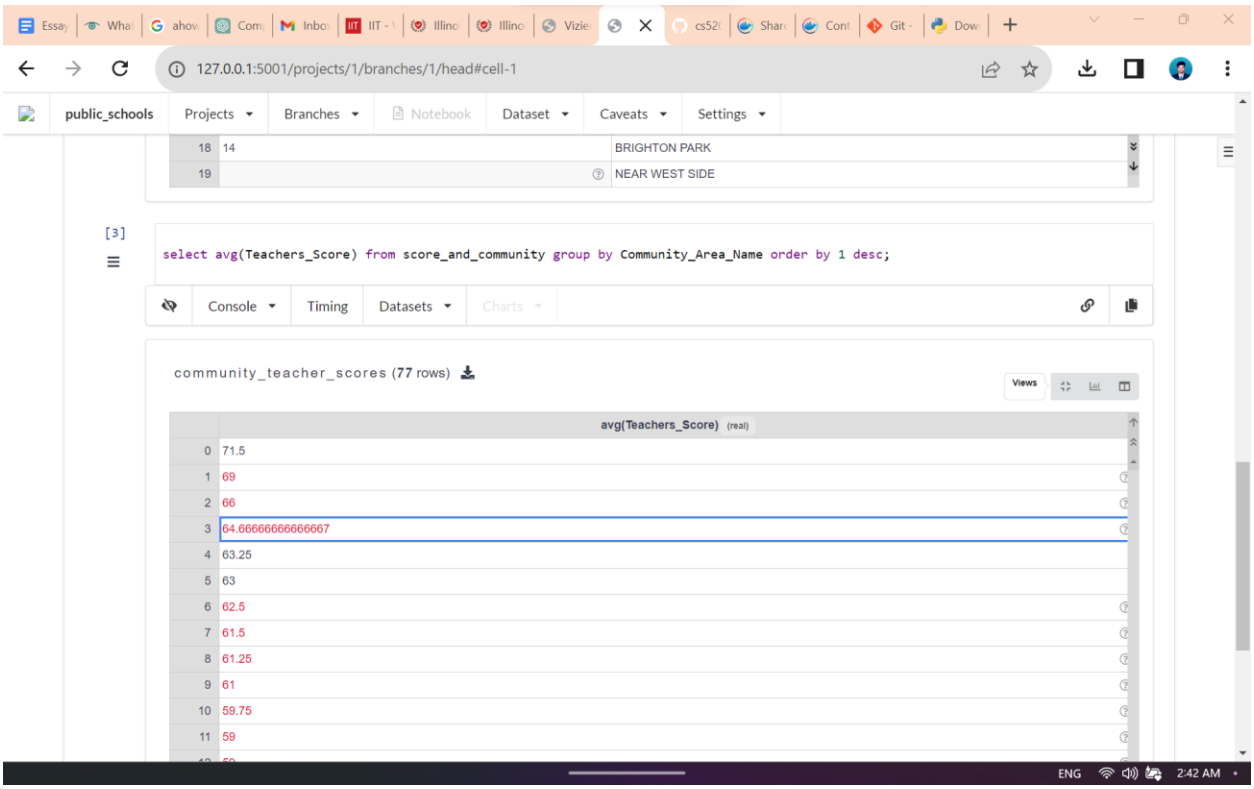
Task 3: Click on one of the question marks for values in the teachers column and take a screenshot.



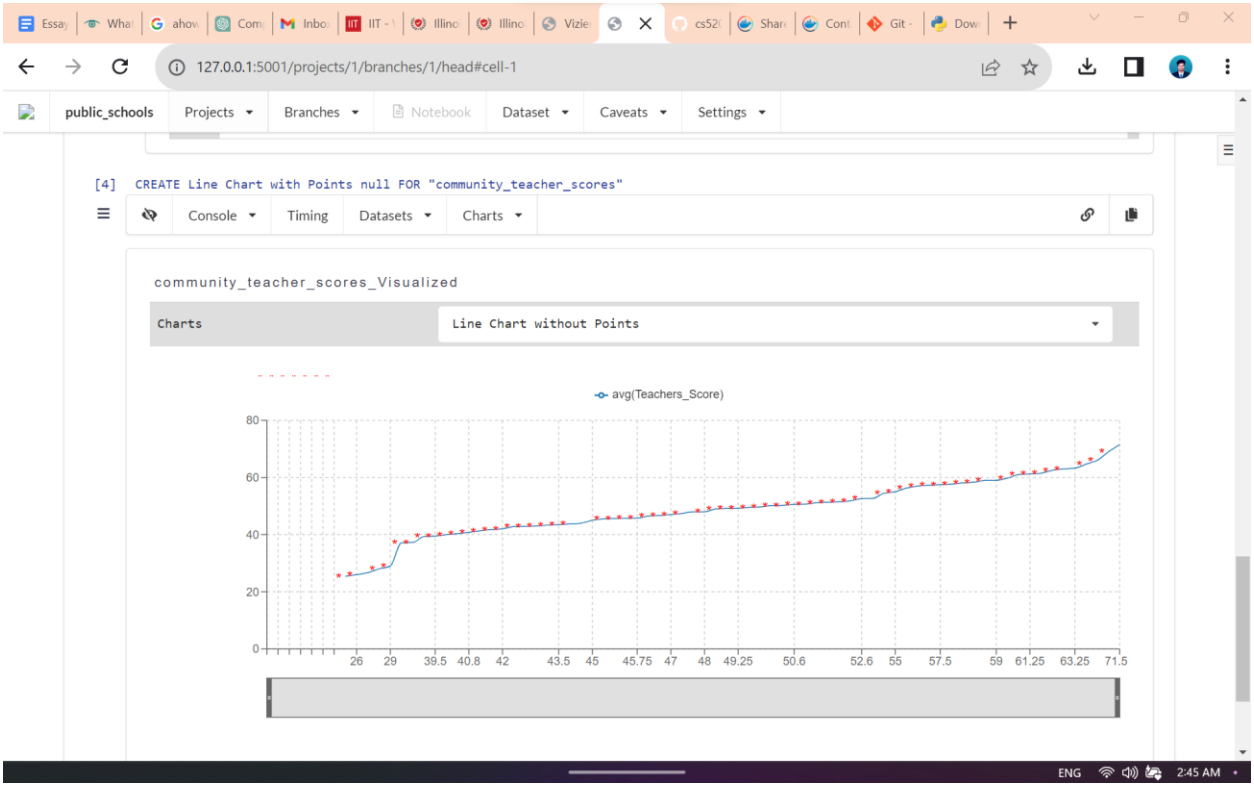
Task 4: Create a SQL cell and write a query that returns columns Teachers_Score and Community_Area_Name. SQL results can be stored as new datasets in Vizier. Call the result dataset score_and_community. And take a screenshot of the result.



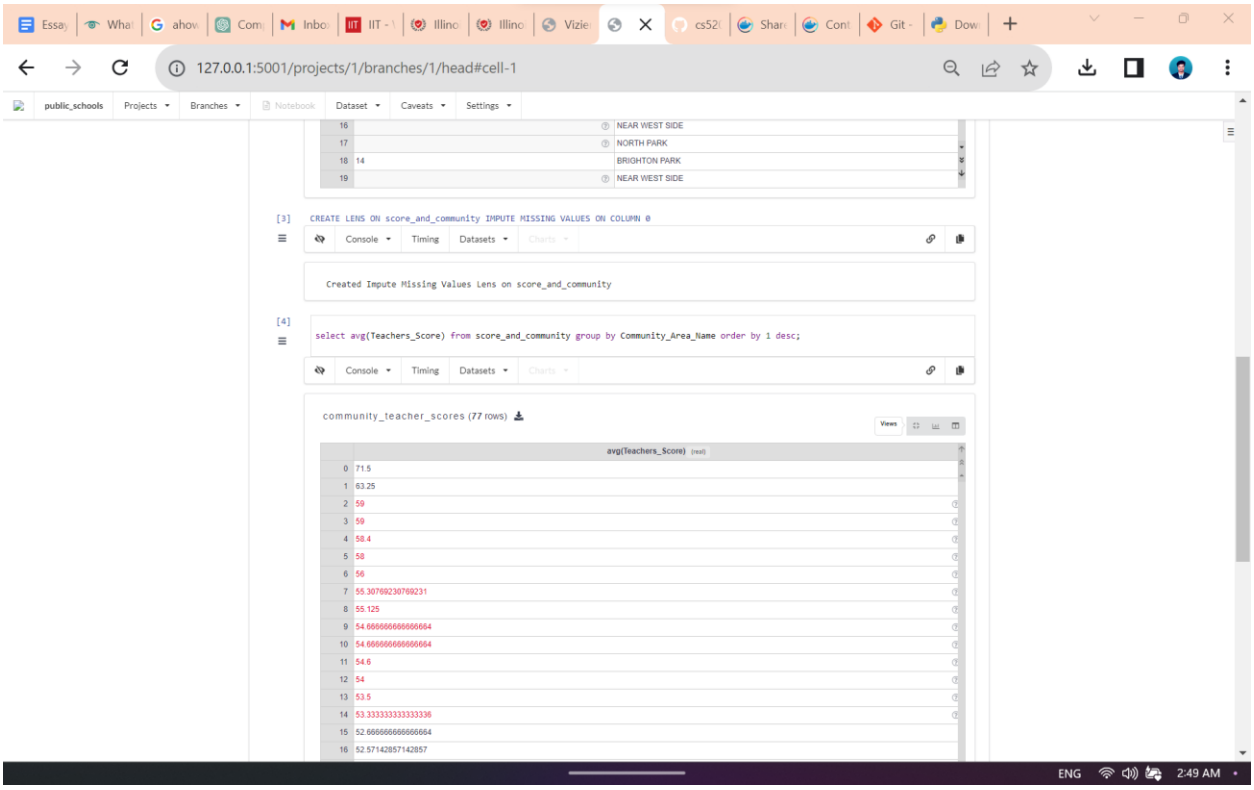
Task 5: Create a SQL cell and write a query over the over the score_and_community dataset that computes the result as described above. Call the result dataset community_teacher_scores. And take a screenshot of the result.

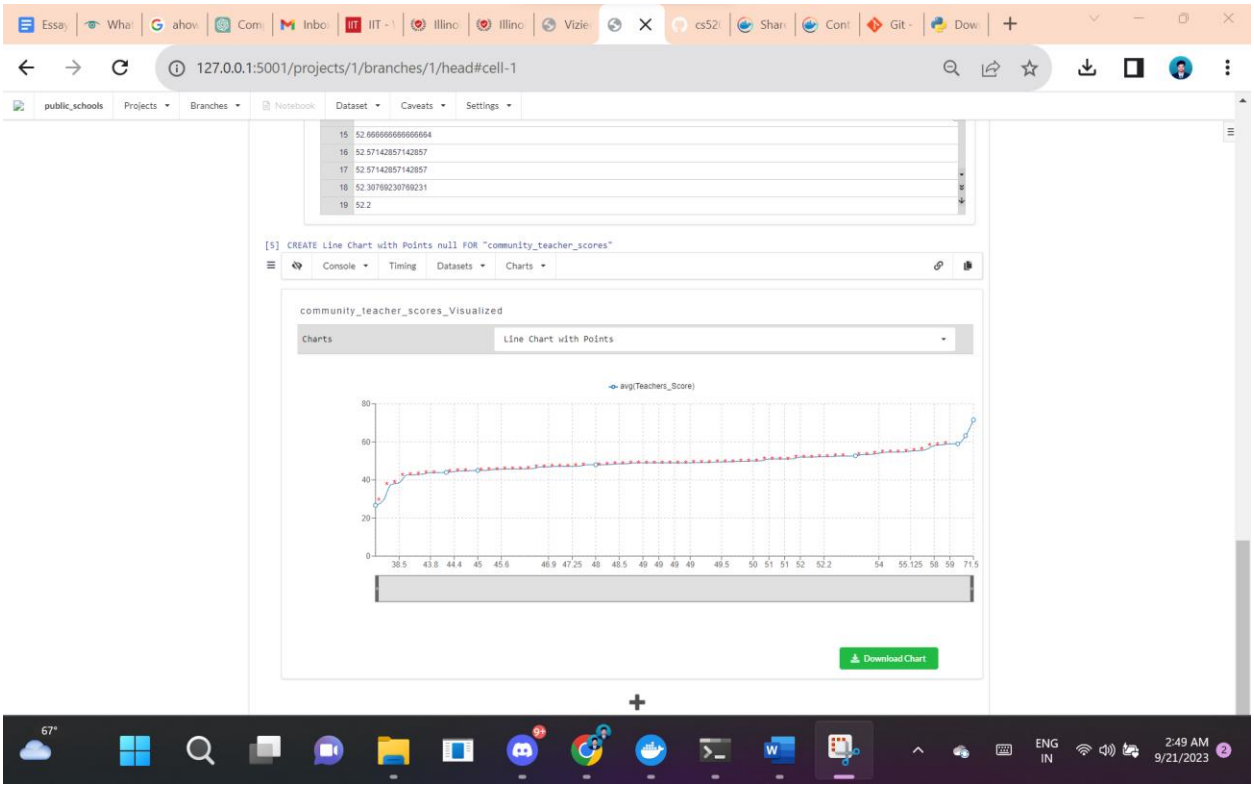


Task 6: Create a line chart of the aggregation result by creating a plot cell and take a screenshot of the result.

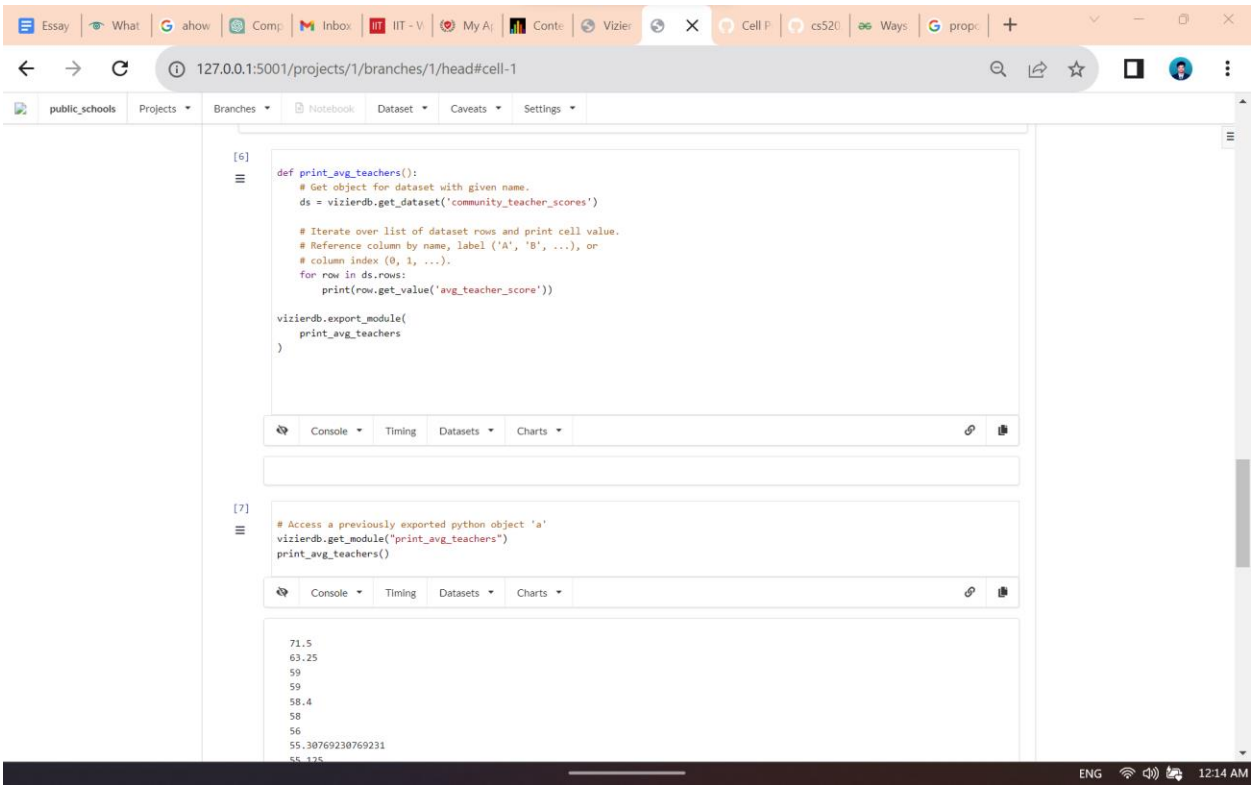


Task 7: Insert a new cell above the SQL cell that computes the average teacher scores (notebooks in Vizier are executed top down) by pressing the three bars below the cell number. Select "Impute Missing Values", select the score_and_community dataset and Teachers_Score as the column to be imputed, and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





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.