

# Data Integration, Warehouse, and Provenance

## Vizier Assignment - 1

### Intro to Vizier

For this assignment we have downloaded docker and downloaded the image.  
Configuration of Coursier and other dependency to use vizier.

Run the following Commands.

- `docker pull iitdbgroup/vizier_iit_cs520_fall23:arm`
- `docker run --name vizier --rm -v `pwd`: /vizier.db -p 5001:5001 -p 8089:8089 iitdbgroup/vizier_iit_cs520_fall23:arm -p 5001`

### Task 1 – Make a new project and load the dataset into it.

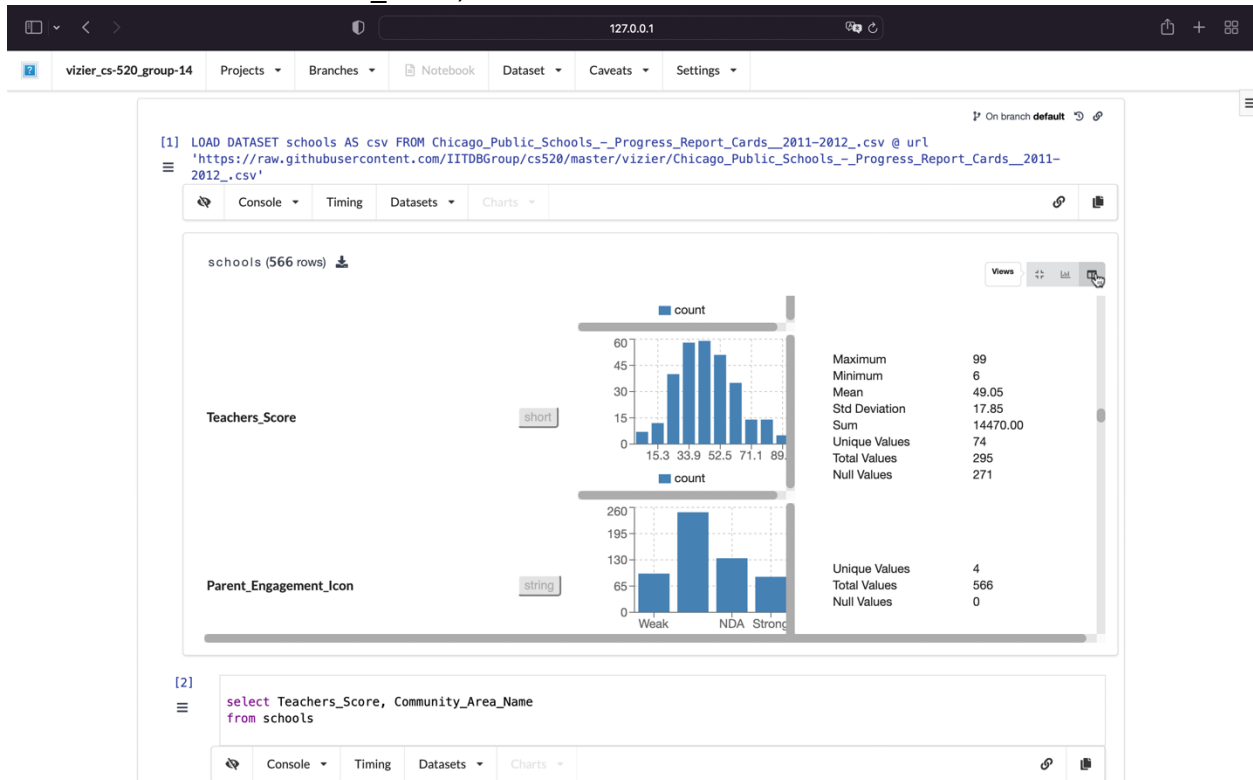
The screenshot shows the Vizier web interface in a browser. The address bar shows the URL `127.0.0.1`. The browser tabs include "LeetCode - The World's Leading Online...", "cs520/README.md at master · IITDBGro...", "First Steps Assignment - CS520 - Data L...", "ITTeaching", and "Vizier DB - vizier\_cs-520\_group-14". The Vizier interface has a top navigation bar with "vizier\_cs-520\_group-14", "Projects", "Branches", "Notebook", "Schools", "Settings", and a "Schools" button. The main content area shows a SQL query in the console:

```
[1] LOAD DATASET schools AS csv FROM Chicago_Public_Schools_-_Progress_Report_Cards_2011-2012_.csv @ url 'https://raw.githubusercontent.com/IITDBGGroup/cs520/master/vizier/Chicago_Public_Schools_-_Progress_Report_Cards_2011-2012_.csv'
```

Below the console, there are tabs for "Console", "Timing", "Datasets", and "Charts". The "Datasets" tab is selected, showing a table of 566 rows. The table has the following columns: School\_ID (int), Name\_of\_School (string), Elementary\_Middle\_or\_High\_School (string), Street\_Address (string), City (string), State (string), and ZIP\_Cod (string). The table displays a list of schools in Chicago, including Charles G Hammond Elementary School, Marvin Camras Elementary School, Eliza Chappell Elementary School, Daniel R Cameron Elementary School, Sir Miles Davis Magnet Elementary Academy, Luther Burbank Elementary School, Lenart Elementary Regional Gifted Center, James N Thorp Elementary School, Walter Payton College Preparatory High School, Roswell B Mason Elementary School, Ira F Aldridge Elementary School, Abraham Lincoln Elementary School, William Penn Elementary School, Christopher Columbus Elementary School, Socorro Sandoval Elementary School, Manley Career Academy High School, Wilma Rudolph Elementary Learning Center, Northside College Preparatory High School, and Frank W Gunsaulus Elementary Scholastic Academy.

School_ID (int)	Name_of_School (string)	Elementary_Middle_or_High_School (string)	Street_Address (string)	City (string)	State (string)	ZIP_Cod (string)
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 60627
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
15	609722	Manley Career Academy High School	HS	2935 W Polk St	Chicago	IL 60612
16	610308	Wilma Rudolph Elementary Learning Center	ES	110 N Paulina St	Chicago	IL 60612
17	609749	Northside College Preparatory High School	HS	5501 N Kedzie Ave	Chicago	IL 60625
18	609958	Frank W Gunsaulus Elementary Scholastic Academy	ES	4420 S Sacramento Ave	Chicago	IL 60632

## Task 2 – Under the Teachers\_Score, select column view to see the distribution.



## Task 3 – Caveats are the ones which has missing value in column.

The screenshot shows the Vizier interface with a dataset named 'schools' (566 rows). The table view is displayed, showing columns: instruction\_score, Leaders\_Icon, Leaders\_Score, Teachers\_Icon, Teachers\_Score, Parent\_Engagement\_Icon, and Parent\_Engagement\_Score. A cell annotation is shown for the 'Teachers\_Score' column, indicating a missing value ('NDA').

**Cell Annotations:**

Comment  
Could not cast 'NDA' to ShortType (in schools)

**Table View (Partial):**

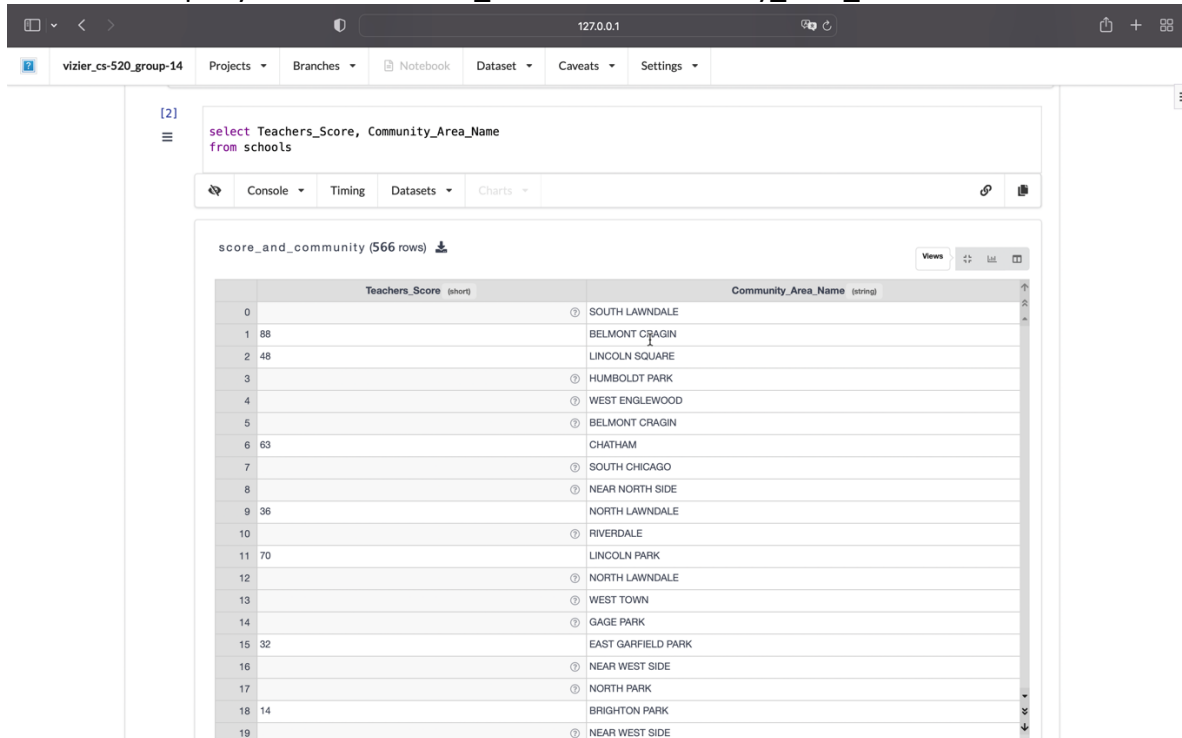
instruction_score	Leaders_Icon	Leaders_Score	Teachers_Icon	Teachers_Score	Parent_Engagement_Icon	Parent_Engagement_Score
43	NDA	NDA	NDA	Weak	43	
41					51	
51					50	
50					46	
54					47	
34					52	
37						
39	NDA	NDA	NDA	NDA		
77	NDA	NDA	NDA	NDA		
58	Average	50	Weak	38	Average	51

**Module Selection:**

Select a module from the list below.

DATA	SQL	VIZUAL
Checkpoint Dataset	SQL Query	Delete Column
Clone Dataset		Delete Row
Declare Parameters		Drop Dataset
Empty Dataset		Filter Columns
Load Dataset		Insert Column
Unload Dataset		Insert Row
Unload File		Move Column

**Task 4 – SQL query to select teacher\_score and community\_area\_number from the dataset.**



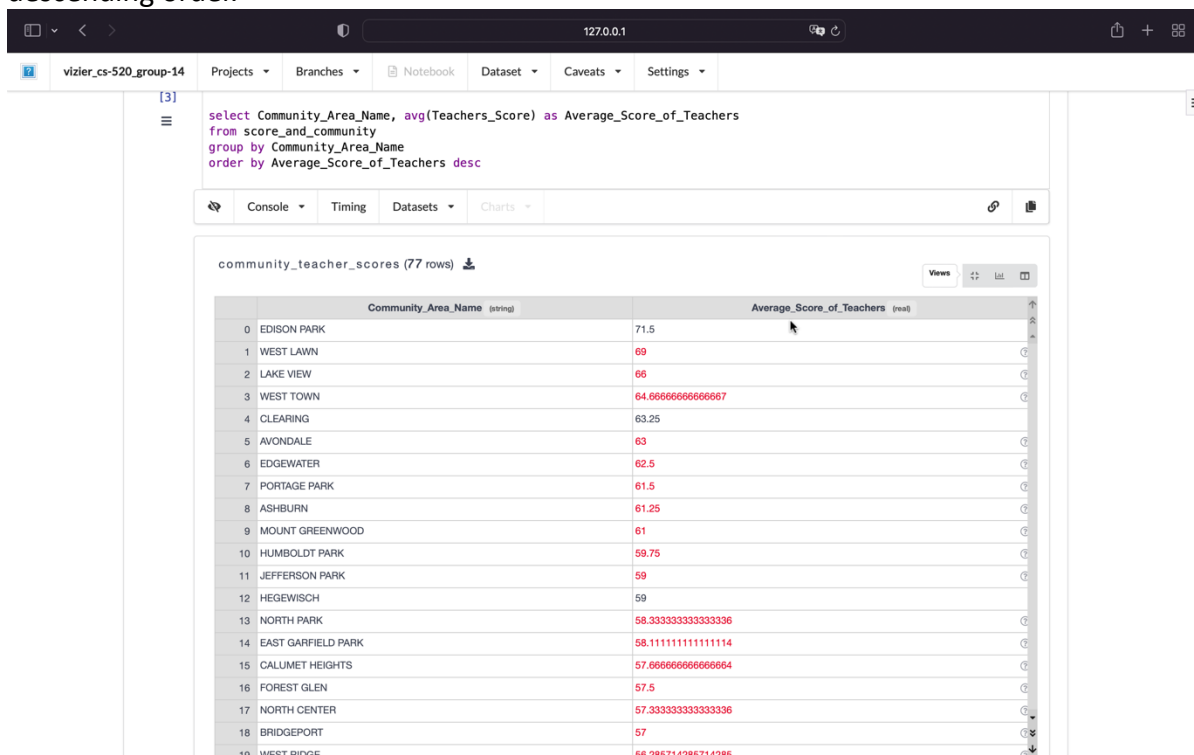
The screenshot shows a Jupyter Notebook interface with a dark theme. The top bar indicates the version is 127.0.0.1. The notebook is titled 'vizier\_cs-520\_group-14'. The active cell contains the following SQL query:

```
[2]  
select Teachers_Score, Community_Area_Name  
from schools
```

Below the query, the results are displayed as a table titled 'score\_and\_community (566 rows)'. The table has two columns: 'Teachers\_Score (short)' and 'Community\_Area\_Name (string)'. The first 20 rows are shown, with the last row truncated with an ellipsis.

	Teachers_Score (short)	Community_Area_Name (string)
0		SOUTH LAWDALE
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 LAWDALE
10		RIVERDALE
11	70	LINCOLN PARK
12		NORTH LAWDALE
13		WEST TOWN
14		GAGE PARK
15	32	EAST GARFIELD PARK
16		NEAR WEST SIDE
17		NORTH PARK
18	14	BRIGHTON PARK
19		NEAR WEST SIDE

**Task 5 – SQL query for dataset score and community which we created in task 4, where we have to select the Community Area Name and Average scores of Teachers where the score is in descending order.**



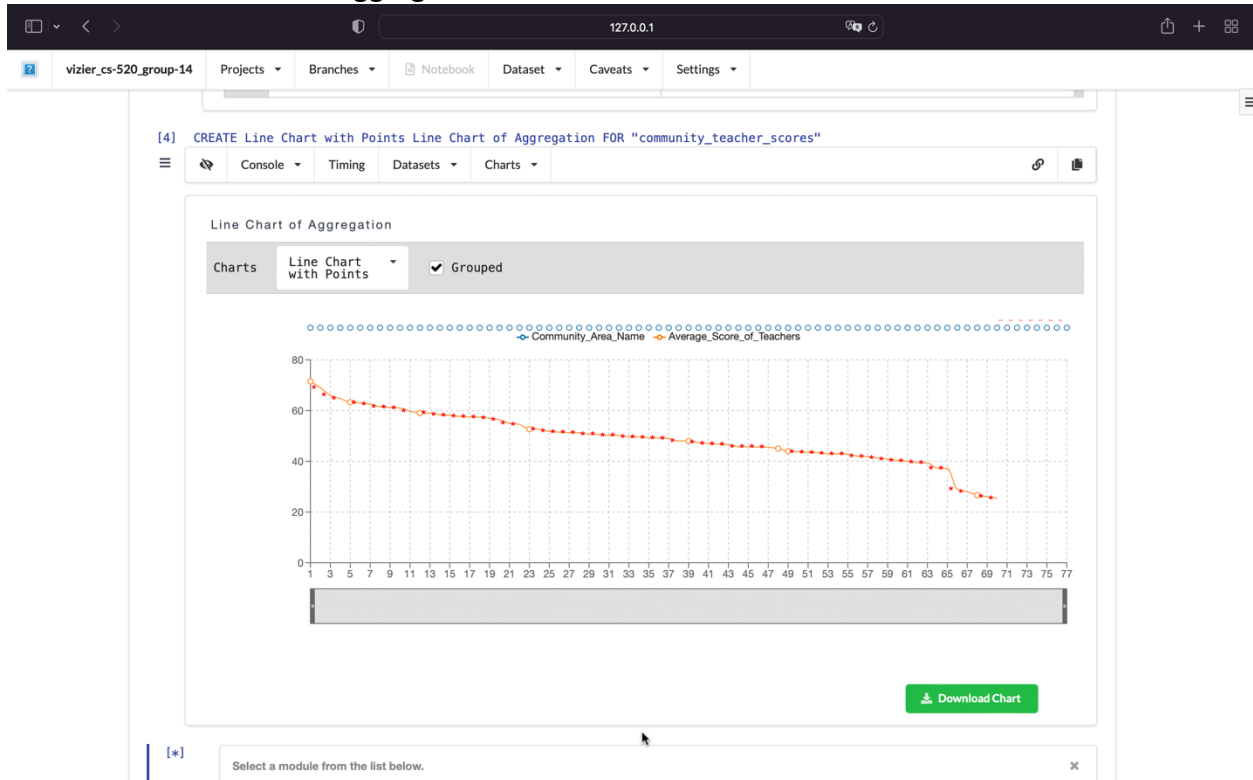
The screenshot shows a Jupyter Notebook interface with a dark theme. The top bar indicates the version is 127.0.0.1. The notebook is titled 'vizier\_cs-520\_group-14'. The active cell contains the following SQL query:

```
[3]  
select Community_Area_Name, avg(Teachers_Score) as Average_Score_of_Teachers  
from score_and_community  
group by Community_Area_Name  
order by Average_Score_of_Teachers desc
```

Below the query, the results are displayed as a table titled 'community\_teacher\_scores (77 rows)'. The table has two columns: 'Community\_Area\_Name (string)' and 'Average\_Score\_of\_Teachers (real)'. The first 20 rows are shown, with the last row truncated with an ellipsis.

	Community_Area_Name (string)	Average_Score_of_Teachers (real)
0	EDISON PARK	71.5
1	WEST LAWN	69
2	LAKE VIEW	66
3	WEST TOWN	64.66666666666667
4	CLEARING	63.25
5	AVONDALE	63
6	EDGEWATER	62.5
7	PORTAGE PARK	61.5
8	ASHBURN	61.25
9	MOUNT GREENWOOD	61
10	HUMBOLDT PARK	59.75
11	JEFFERSON PARK	59
12	HEGEWISCH	59
13	NORTH PARK	58.33333333333333
14	EAST GARFIELD PARK	58.11111111111111
15	CALUMET HEIGHTS	57.66666666666666
16	FOREST GLEN	57.5
17	NORTH CENTER	57.33333333333333
18	BRIDGEPORT	57
19	WEST DIVIDE	56.28571428571428

## Task 6 – Line chart of the aggregation which we made in above task.



## Task 7 – Add a cell above the Average Scores of Teachers and add Impute Missing Values and select mean, so it will change in below steps which we performed.

Community Area	Average Score of Teachers
RIVERDALE	70
LINCOLN PARK	
NORTH LAWNDALE	
WEST TOWN	
GAGE PARK	
EAST GARFIELD PARK	32
NEAR WEST SIDE	
NORTH PARK	
BRIGHTON PARK	14
NEAR WEST SIDE	

[3] Impute Missing Values

Dataset: score\_and\_community

Columns:

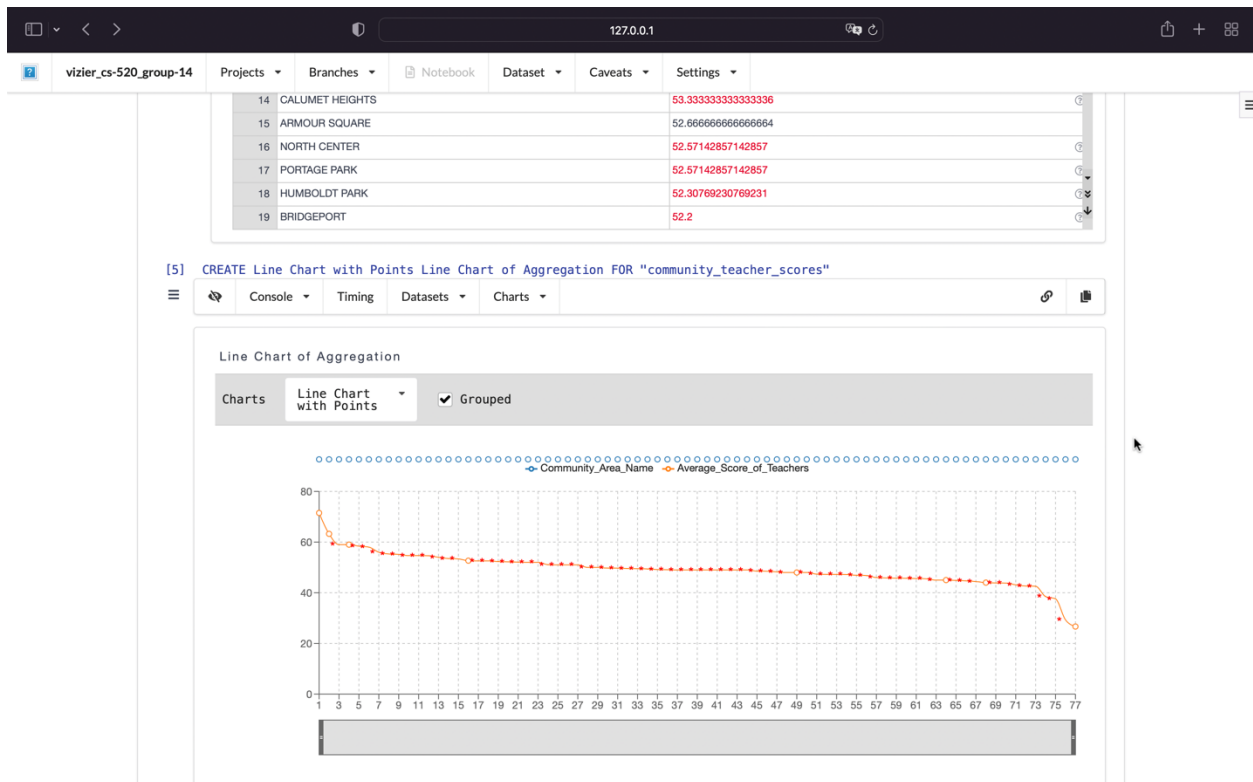
Column	Model
Teachers_Score	Mean
<Pick One For Me>	<Pick One For Me>

Change Command Dismiss Submit

Created Impute Missing Values Lens on score\_and\_community

[4]

```
select Community_Area_Name, avg(Teachers_Score) as Average_Score_of_Teachers
from score_and_community
group by Community_Area_Name
```



**Task 8** – Create a Python cell to print all the rows of average scores of teacher using vizierdb.export\_module. Create another cell and print the data.

```
[6]
def print_avg_teachers(ds):
    for row in ds.rows:
        print(row.get_value('Average_Score_of_Teachers'))
vizierdb.export_module(print_avg_teachers)

[7]
ds = vizierdb.get_dataset('community_teacher_scores')
vizierdb.get_module("print_avg_teachers")
print_avg_teachers(ds)
```

```
71.5
63.25
59
59
58.4
58
56
55.30769230769231
55.125
54.66666666666666
54.66666666666666
54.6
54
53.5
53.33333333333333
52.66666666666666
52.57142857142857
52.57142857142857
52.30769230769231
```

**Task 9** – Python cell to filter the rows where the average score of teachers should be more than 30.0.

The screenshot shows a Jupyter Notebook interface with a dark theme. The top bar includes a browser address bar with '127.0.0.1' and navigation icons. Below the top bar is a toolbar with tabs for 'vizier\_cs-520\_group-14', 'Projects', 'Branches', 'Notebook', 'Dataset', 'Caveats', and 'Settings'. The main area contains a list of scores and a code cell.

46  
45.84615384615385  
45.76470588235294  
45.714285714285715  
45.6  
45.5  
45  
44.857142857142854  
44.69230769230769  
44.4  
44  
43.857142857142854  
43.8  
43.125  
42.666666666666664  
42.54545454545455  
38.5  
37.57142857142857  
29.333333333333332  
26.666666666666668

```
[8]: df = vizierdb.get_data_frame('community_teacher_scores')  
print(df[df.Average_Score_of_Teachers < 30.0])
```

Community_Area_Name	Average_Score_of_Teachers
75 ROGERS PARK	29.333333
76 AVALON PARK	26.666667

Connected to vizier @ http://localhost:5001/vizier-db/api/v1/