Assignment #2: Spatial Database

Toronto Metropolitan University

Shakeeb Tahir

500837388

Dr. Christopher Daniel

SA8902 – Database Management and Spatial Technologies

Table of Contents:

1. Introduction....3
1.1 Background....3
1.2 Research Questions....4
2. Data Modelling and Database Design....4
2.1 Data Sources....4
2.2 Modelling and Design....5
3. ER Diagram....6
4. Data Dictionary....8
4.1 Entity Instance Descriptions....12
4.2 Relationships Table....14
5. Spatial Data Type....14
6. Geography and Geometry Data Types....14
7. Spatial Indexes....15
8. SQL Statements....16
9. Conclusion....20

1. Introduction

1.1 Background

Access to after school resources such as libraries or recreation facilities could possibly have a large impact on a students performance at school. This is because these are places where students can go after school to relax, de-stress, and continue learning. By having access to these resources in their neighbourhoods it is more likely to foster not only intellectual growth but also overall well-being. Having a proximity to libraries provides students with a rich resource hub to broaden their knowledge base. Many libraries also often have partnerships with local schools which can help to facilitate a seamless integration of educational resources into the students' daily lives. Similarly, access to nearby recreational facilities can contribute significantly to a student's physical and mental state, as they offer a space for rejuvenation and social interaction. When schools are in proximity to such resources, students could be more likely to perform better academically, as they have more resources to learn from while also being able to take care of their mental and physical wellbeing.

As such, for this project I have built a database of schools within each Toronto neighbourhood, as well as all the recreational facilities and libraries that are in each neighbourhood. This will provide students with the ability to know where resources are available near their school and what type of services they offer. Furthermore, in order to see if there is a relationship between the access students have to these resources (ie. being able to travel to these places after school) and how well students do at school, I have also included a table that tracks the Ontario Secondary School Literacy Test (OSSLT) pass rate in each neighbourhood as well as the high school graduation rate. Unfortunately, I was only able to find this data related to the Toronto Catholic District School Board (TCDSB) which is why this database only contains Secondary Schools within the TCDSB.

1.2 Research Questions

The research questions the database intends to help answer are:

1) What is the distance from each school to the nearest library?

2) For neighbourhoods with a <60% OSSLT pass rate, do they have any schools and if so what is the distance from those schools to the nearest library?

3) Which schools have recreational facilities within a 1 km buffer and how many?

4) In neighbourhoods with both schools and recreational facilities, what is the average distance from the school to a recreational facility?

5) Which schools have more than 1 library in a 1 km buffer distance?

2. Data Modelling and Database Design

2.1 Data Sources

Data Name	Description	Format	Source
Neighbourhoods	Shapefile containing polygons of Toronto neighbourhoods	Shapefile	Toronto Open Data Portal
TCDSB Schools	Shapefile containing points of TCDSB schools	Shapefile	Toronto Open Data Portal
Libraries	CSV file containing general information and lat/long coordinates of Toronto Public Library branches	CSV	Toronto Open Data Portal
Wellbeing Toronto - Education	Excel file containing education statistics of Toronto neighbourhoods (% passing OSSLT and high school graduation rates)	XLSX	Toronto Open Data Portal
Wellbeing Youth - Recreation	Shapefile containing points of cultural, arts, sports, and recreation program locations geared to youth and young adults, including drop-in programs and independent community centres	Shapefile	Toronto Open Data Portal

2.2 Modelling and Design

Once all the data sets were downloaded, they were first loaded into ArcGIS pro to modify and clean so that we are only left with the data we need for the database. For the neighbourhoods shapefile, all the fields were removed except for the ID field and neighbourhood name. For the TCDSB shapefile, it was filtered by removing the points which are not Secondary Schools and then removing any unnecessary fields like municipality or status. For the library CSV file, it was turned into a shapefile by converting the latitude and longitude coordinates into points. The Wellbeing Toronto Education excel file which contains data about TCDSB neighbourhoods and their OSSLT pass rate and high school graduation rates were just cleaned by removing unnecessary fields and renaming the fields we need to be shorter. The Wellbeing Youth Recreation shapefile just removed unnecessary fields like longitude and latitude, municipality, etc.

All the point shapefiles (schools, libraries, recreation facilities) were then given a NeighbourhoodID column and each point in the shapefiles were given the value of the neighbourhood they were in. This was done by using ArcGIS Pro to find the intersection of the point shapefile and the neighbourhood shapefile. A NeighbourhoodID column was needed since we needed a way to link the points to whichever neighbourhood they belong to and so that we can connect the data together. We did not need to “Normalize” the data as it was already done so, since all these files were made separately instead of being in one large table. The NeighbourhoodID will be used as the foreign key to connect the tables to one another and separate unique id columns (primary keys) like “SchoolID”, “LibraryID”, and “RecreationFacilityID” were also created in order to uniquely identify each row in the shapefile data sets. The wellbeing education excel data and the neighbourhood shapefile already had existing “NeighbourhoodID” fields which matched so we did not need to create any new fields to link them to each other.

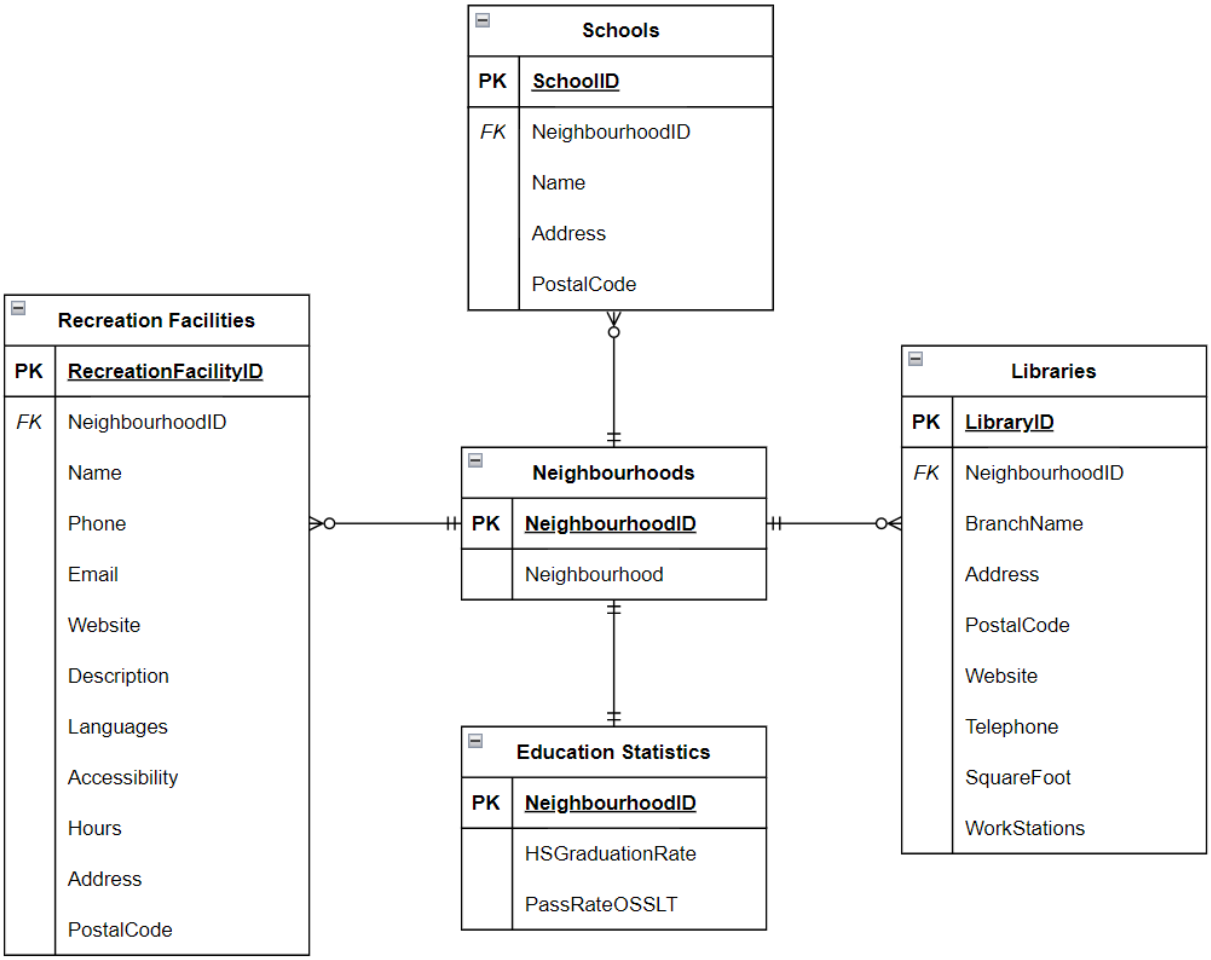
Now that the data was cleaned, and we had a way to uniquely identify each row as well as a way to connect the data to each other, ArcCatalog was used to import all the data into Microsoft SQL Server Management Studio.

By using a database, we also avoid a lot of modification problems that would traditionally occur if we just had all our data in an excel sheet. With a database we do not have to worry about inserting new data as adding new data such as adding a new library would not impact any of the other tables. Meanwhile if we were to try and do this in an excel sheet, we could insert the wrong type of data into it like accidentally inputting the name into the ID column. A database prevents us from doing this as we need to input the correct data type. The same problem is avoided in a database when updating

data. In terms of deletion, in an excel file we could accidentally delete data that is required like a neighbourhood ID or a unique ID in a row, but this would not be allowed in the database as those fields would not be allowed to be null.

3. ER Diagram

For the ER Diagram, the fields added when importing the data into MS SQL Server (ObjectID, Shape) are not included as it makes the diagram easier to understand



3.1 ER Diagram Explanation and Discussion

The neighbourhoods and education statistics contain the same primary key of NeighbourhoodID as those are the primary ways to identify the rest of the data in the tables. This is also how they are connected as knowing the NeighbourhoodID allows us to know which neighbourhood it refers to and what the high school graduation rate and OSSLT first attempt pass rates of the neighbourhood are. Both the Neighbourhoods and Education Statistics have a mandatory-one relationship to each other as each neighbourhood can only have one education statistic and each education statistic can only refer to one neighbourhood. They do not have an optional (zero) relationship as there needs to be an education statistic associated with each neighbourhood and vice versa. For the Neighbourhood tables relationship to each of the points files (libraries, schools, recreational facilities) they are all optional-many as a neighbourhood can have 0 or an infinite amount of schools, libraries, recreation facilities. It is not required that a neighbourhood have these resources but at the same time it could have many of them. For each of the point files (libraries, schools, recreational facilities) relationships to Neighbourhoods they have a mandatory-one relationship as a point can only belong to a minimum and maximum of one neighbourhood. The point needs to exist in one of the neighbourhoods and it can only be in one as it cannot exist in two neighbourhoods at once. While theoretically a point could be in the middle of two neighbourhood boundaries, this is not the case in the data we have and administrations do not usually place buildings where they would overlap into different neighbourhoods so we can safely assume that a point can only exist in one neighbourhood and not many.

4. Data Dictionary

NEIGHBOURHOODS:

Column Name	Data Type (Length)	Key	Null Status	Default Value	Remarks	Description
NeighbourhoodID	int	Primary Key	Not Null	None	These IDs were automatically assigned from the downloaded data	Unique ID for each neighbourhood
Neighbourhood	nvarchar(80)	No	Not Null	None	80 characters is generally enough for a neighbourhood name	Name of the neighbourhood

EDUCATION STATISTICS:

Column Name	Data Type (Length)	Key	Null Status	Default Value	Remarks	Description
NeighbourhoodID	int	Primary Key	Not Null	None	These IDs were automatically assigned from the downloaded data	Unique ID for each neighbourhood
HSGraduationRate	numeric(3,2)	No	Not Null	None	Ranges from 0-1 (E.g 0.4 indicates 40%) Numeric(3,2) gives 3 digits, 2 of which are decimal points. This is enough to hold the range	Percent of students graduating high school
PassRateOSSLT	numeric(5,2)	No	Not Null	None	Ranges from 0-100 (E.g 40 indicates 40%) Numeric(5,2) gives 5 digits, 2 of which are decimal points. This is enough to hold the range	Pass rate of the OSSLT on the first attempt

SCHOOLS:

Column Name	Data Type (Length)	Key	Null Status	Default Value	Remarks	Description
SchoolID	int	Primary Key	Not Null	DBMS Supplied	Initial Value = 1; Increment = 1	Unique ID for each school
NeighbourhoodID	int	Foreign Key	Not Null	None	REF: NEIGHBOURHOODS	Unique ID of the neighbourhood
Name	nvarchar(80)	No	Not Null	None	80 characters is generally enough for a name	Name of the school
Address	nvarchar(80)	No	Not Null	None	80 characters is generally enough for an address	Address of the school
PostalCode	char(7)	No	Not Null	None	7 characters for a postal code including space is enough	Postal Code of the school

LIBRARIES:

Column Name	Data Type (Length)	Key	Null Status	Default Value	Remarks	Description
LibraryID	Int	Primary Key	Not Null	DBMS Supplised	Initial Value = 1; Increment = 1	Unique ID for each library
NeighbourhoodID	int	Foreign Key	Not Null	None	REF: NEIGHBOURHOODS	Unique ID of the neighbourhood
BranchName	nvarchar(80)	No	Not Null	None	80 characters is generally enough for a name	Name of the library
Address	nvarchar(80)	No	Not Null	None	80 characters is generally enough for an address	Address of the library
PostalCode	char(7)	No	Not Null	None	7 characters for a postal code including space is enough	Postal code of the library
Website	nvarchar(50)	No	Null	None	50 characters is generally enough for an address	Website of the library
Telephone	nvarchar(12)	No	Null	None	12 characters (Format: xxx-xxx-xxxx) is generally enough for a phone number	Telephone number of the library
SquareFoot	int	No	Null	None		Square Footage of the library
WorkStations	int	No	Null	None		Number of workstations library has

RECREATION FACILITIES

Column Name	Data Type (Length)	Key	Null Status	Default Value	Remarks	Description
RecreationFacilityID	int	Primary Key	Not Null	None	Initial Value = 1; Increment = 1	Unique ID for each recreation facility
NeighbourhoodID	int	Foreign Key	Not Null	None	REF: NEIGHBOURHOODS	Unique ID of the neighbourhood
Name	nvarchar(80)	No	Not Null	None	80 characters is generally enough for a name	Name of the recreation facility
Phone	nvarchar(12)	No	Null	None	12 characters (Format: xxx-xxx-xxxx) is generally enough for a phone number	Phone number of the recreation facility
Email	nvarchar(50)	No	Null	None	50 characters is generally enough for an email	Email of the recreation facility
Website	nvarchar(50)	No	Null	None	50 characters is generally enough for a website	Website of the recreation facility
Description	nvarchar(MAX)	No	Null	None	The description can be as long as possible so it is set to max	Description about what services and resources are provided
Languages	nvarchar(MAX)	No	Null	None	The languages supported and their description can be as long as possible so it is set to max	Languages that are supported at the facility
Accessibility	nvarchar(MAX)	No	Null	None	The accessibility support and its	Description of accessibility

					description can be as long as possible so it is set to max	features of the facility
Hours	nvarchar(254)	No	Null	None	254 should be enough to list the hours of operation	Description of the hours of the week they are open
Address	nvarchar(80)	No	Not Null	None	80 characters is generally enough for an address	Address of the facility
PostalCode	char(7)	No	Not Null	None	7 characters for a postal code including space is enough	Postal code of the facility

4.1 Entity Instance Descriptions

Neighbourhoods:
The neighbourhoods entity holds data about each of the neighbourhoods in the city of Toronto. It contains 140 neighbourhoods (the historical neighbourhoods before the recent update) and each neighbourhood is associated with a NeighbourhoodID which was already assigned by the creator of the shapefile (Toronto Open Data Portal). This NeighbourhoodID is the primary key of the table as it is used to uniquely identify each neighbourhood. Both the NeighbourhoodID and Neighbourhood also can't be null as we need data about them in order for the database to function.

Education Statistics:
The education statistics entity holds data about each neighbourhood's high school graduation rate and Ontario Secondary School Literacy Test (OSSLT) first attempt pass rate. The primary key in this case is the NeighborhoodID as it is used to identify the neighbourhoods and no foreign key is required as the id and its values refer to the same neighbourhoods in the Neighbourhood table. None of the data can be null as we need the education information about each neighbourhood.

Schools:
The school entity contains information about all the high schools in the Toronto Catholic District School Board (TCDSB). The primary key in this case is a SchoolID that was made which is used to uniquely identify each school. The foreign key is the NeighbourhoodID as we can use it to link each school to whatever neighbourhood it belongs to. None of the data is allowed to be null as it is all information that is required for a school to exist as we cannot have a school without a postal code, name, address, etc.

Libraries:
The library entity contains information about all of the libraries within Toronto neighbourhoods. This includes useful information a person would want to know like how many workstations it has, their website, phone number, etc. The primary key created for this table was the LibraryID which is used to uniquely identify each of the libraries. The foreign key is once again the NeighbourhoodID as we need a way to link the libraries to whichever neighbourhood they belong to. Only the address data and the id's are not allowed to be null as they need to exist in order for the library to exist. However, fields like the website, telephone, etc, can be null as we may not know this information at the time or it could not currently exist.

Recreation Facilities:
The recreation facilities entity contains information about all of the recreation facilities within Toronto neighbourhoods. These recreation facilities offer culture, arts, and recreation programs geared towards young adults. The primary key is the RecreationFacilitiesID which is used to uniquely identify each recreation facility we have. The foreign key again is the NeighbourhoodID which links the recreation facility to whichever neighbourhood it belongs to. Similar to the Library entity, only the id fields and the address fields (name, address, postal code) are not allowed to be null as they are required information in order for a recreation facility to exist. Other fields like the website, description, hours, etc, can be null as they are not needed for a recreation facility to exist and they do not need to exist themselves as we may not know the hours for the recreation facility or it may not have a website, etc.

4.2 Relationships Table

Parent	Child	Referential Integrity Constraint	On Update	On Delete
Neighbourhoods	Libraries	NeighbourhoodID in Libraries must exist in Neighbourhoods	No	No
Neighbourhhods	Schools	NeighbourhoodID in Schools must exist in Neighbourhoods	No	No
Neighbourhoods	Recreation Facilities	NeighbourhoodID in RecreationFacilities must exist in Neighbourhoods	No	No

5. Spatial Data Type

The spatial data type used was the geometry type. The geometry type was used since the study area of the project was Toronto and as such it does not have a large geography, so using the geography type is not really necessary. The geography type is more complicated as it is 3 dimensional which makes it more computationally expensive and would be better suited if the database had a study area like an entire country rather than a city. By using the geometry data type, it is more accurate as it uses a flat plane to perform calculations when using spatial functions and is also less computationally expensive.

6. Geography and Geometry Data Types

The geography data type stores geodetic information that takes into account the curvature of the earth and makes use of a geographic coordinate system. It is 3 dimensional as it uses latitude and longitude coordinates to perform calculations. This makes it slightly more inaccurate as the calculations are more complex and also more computationally expensive. The geometry data type on the other hand treats spatial data as lying flat on a plane. This makes the calculations more accurate and less computationally expensive. It is best suited to be storing data in a projected coordinate system.

7. Spatial Indexes

Spatial indexes in relational databases help to optimize the retrieval and query performance of spatial data. In order to use spatial indexes in a relational database we need the table to have a clustered primary key and spatial indexes can only be created for geography or geometry data types. When spatial queries are executed in a relational database like SQL server, they go through two filters. The first filter is called the primary filter which is a fast approximation of where the objects are located and then it runs a second filter if a spatial function (such as STWithin()) is used which is a more accurate but computationally expensive process which refines the initial objects that were selected from the first filter and removes any false positives which may have occurred. The way it works is by creating a defined grid (such as 4x4 or 16x6) and then checking which spaces in the grid have something within them. So if we were to use a spatial function such as STWithin() to see if a point falls within a polygon, the grid is checked to see if the point is in an empty grid cell and if so then it does not need to run the second part of the filter which would do the computationally expensive calculations. Spatial indexes with a 16x16 grid were added to each of the spatial data tables (neighbourhoods, libraries, schools, recreational facilities) in order to speed up the performance of any queries that are made since there files all have geographic data that is spatially represented and needs intensive calculations to be done on it.

8. SQL Statements

1) What is the distance from each school to the nearest library?

```
WITH RankedDistances AS (
SELECT
SCHOOLS.Name AS SchoolName,
LIBRARIES.BranchName AS NearestLibrary,
SCHOOLS.Shape.STDistance(LIBRARIES.Shape) * 111300 AS DistanceToLibrary,
ROW_NUMBER() OVER (PARTITION BY SCHOOLS.Name ORDER BY
SCHOOLS.Shape.STDistance(LIBRARIES.Shape)) AS RowNum

FROM SCHOOLS
CROSS JOIN LIBRARIES
)

SELECT
SchoolName,
NearestLibrary,
DistanceToLibrary
FROM RankedDistances
WHERE RowNum = 1
ORDER BY SchoolName;
```

The SQL query creates a CTE in which the distance to the libraries from each school are ranked and from the ranked distances it selects the highest ranked library which represents the shortest distance to the school. The distances are multiplied by 111300 as all of our shapefile data was using the SRID of 4326 (WGS 84 projection) which returns measurements in decimal degrees. To convert the measurement from decimal degrees to metres it needs to be multiplied by 111300 (Source: https://gis.stackexchange.com/questions/211038/how-does-qgis-calculate-distances-when-the-data-is-not-projected-wgs84#:~:text=1%20degree%20%3D%20111300%20metres) This helps answer the geographic question of which library is the nearest to each school as we now have a list of each school, the closest library, and the distance to that library. This can be used to find out which library students from each school should go to as it is the closest.

2) For neighbourhoods with a <60% OSSLT pass rate, do they have any schools and if so what is the distance from those schools to the nearest library?

```
SELECT
NEIGHBOURHOODS.Neighbourhood,
EDUCATIONSTATISTICS.PassRateOSSLT,
SCHOOLS.Name AS SchoolName,
LIBRARIES.BranchName,
SCHOOLS.Shape.STDistance(LIBRARIES.Shape) * 111300 AS DistanceToLibrary

FROM EDUCATIONSTATISTICS
INNER JOIN NEIGHBOURHOODS ON NEIGHBOURHOODS.NeighbourhoodID =
EDUCATIONSTATISTICS.NeighbourhoodID
LEFT JOIN SCHOOLS ON NEIGHBOURHOODS.NeighbourhoodID =
SCHOOLS.NeighbourhoodID
LEFT JOIN LIBRARIES ON NEIGHBOURHOODS.NeighbourhoodID =
LIBRARIES.NeighbourhoodID

WHERE EDUCATIONSTATISTICS.PassRateOSSLT < 60;
```

This SQL query joins all the tables together and checks which rows have a pass rate for the OSSLT < 60%. From these rows, the distance is then calculated to the nearest library. The distance is multiplied by 111300 to convert it from decimal degrees to metres. LEFT JOIN is used for the schools and libraries since we want to see which neighbourhoods do not have any schools or libraries in them. This answers the geographic question of knowing if the neighbourhoods that have <60% OSSLT scores have schools in them as it shows the name of the school or NULL and tells us the distance to the nearest library from that school. This is important because it could potentially show that a library is being underfunded within an area or if schools should be built in certain neighbourhoods.

3) Which schools have recreational facilities within a 1 km buffer and how many?

```
SELECT
SCHOOLS.Name AS SchoolName,
COUNT(RECREATIONFACILITIES.Shape) AS RecreationFacilityCount

FROM SCHOOLS, RECREATIONFACILITIES

WHERE
SCHOOLS.Shape.STBuffer(0.009).STIntersects(RECREATIONFACILITIES.Shape) = 1

GROUP BY SCHOOLS.Name;
```

This query creates a 1 km buffer around every school school and checks to see if the buffer intersects with any recreation facilities. 0.009 represents 1 km in decimal degrees (~1000 / 1113000) and =1 checks to see if the intersection was true. We use 0.009 since the shapefiles all had projected coordinate systems of WGS 84 (SRID 4326) which uses decimal degrees. The number of recreation facilities that were in the buffer of each school is then listed. This helps us answer the geographic question of knowing how many recreation facilities are near each school and can help guide us to know which schools need more recreation facilities built near them.

4) In neighbourhoods with both schools and recreational facilities, what is the average distance from the school to a recreational facility?

```
SELECT
SCHOOLS.Name AS SchoolName,
AVG(SCHOOLS.Shape.STDistance(RECREATIONFACILITIES.Shape)) * 111300 AS
AvgDistanceToRecreationFacility

FROM SCHOOLS
INNER JOIN
RECREATIONFACILITIES ON RECREATIONFACILITIES.NeighbourhoodID =
SCHOOLS.NeighbourhoodID
GROUP BY
SCHOOLS.Name
ORDER BY
AvgDistanceToRecreationFacility;
```

This SQL query joins the Schools table to the RecreationFacilities using an inner join so only neighbourhoods which have both of them are selected. The average distance from the school to the recreational facilities is then calculated and ordered from least to greatest. This query helps answer the geographic question of knowing what the average distance to travel from a school to a recreational facility is and can help us know which schools have recreational facilities that are too far away on average in order to prioritize providing those schools with additional facilities.

5) Which schools have more than 1 library in a 1 km buffer distance?

```
WITH SchoolLibraries AS (
SELECT
SCHOOLS.Name AS SchoolName,
COUNT(LIBRARIES.Shape) AS LibraryCount
FROM SCHOOLS,  LIBRARIES
WHERE
SCHOOLS.Shape.STBuffer(0.009).STIntersects(LIBRARIES.Shape) = 1
GROUP BY SCHOOLS.Name
)

SELECT
SchoolName,
LibraryCount
FROM SchoolLibraries
WHERE LibraryCount > 1;
```

This query creates a CTE which selects the school and the number of libraries it has within a 1 km buffer. The schools with more than 1 library are then selected. This helps to answer the geographic question of knowing which schools might already have the most resources and that they might not need as many resources like libraries built since they already have more than 1 in such a close distance.

9. Conclusion

In conclusion, we can see how the database we have built helps us to answer a variety of different geographic questions. Not only are we able to answer geographical questions but we also have a useful resource for storing data related to resources that can be accessed after school. The database could be useful in helping to guide future policy making decisions to government officials or city planners as they can gain a better understanding of which neighbourhoods or schools are lacking resources and where they should prioritize resources in order to ensure equal access of opportunities for all individuals and not just those living in well off neighbourhoods. This is important as the proximity of libraries and recreational facilities to schools could possibly increase the overall development and academic success of students. Having access to libraries can help students continue their learning after school while being in a safe and welcoming environment. Likewise, having nearby recreational facilities can contribute to their physical and mental well-being, making them more likely to perform better at school.