MSDA 3040/MSIT 3860 - FUNDAMENTALS DATA ENGINEERING

Project Title: Recreational Spring Open Gym Activities

GROUP NUMBER: 5

NAMES and CLARK IDs: Kanishma Subbiah - C70293277

Project Title: Recreational Spring Open Gym Activities

OVERVIEW:

The data is about **Recreational Spring Open Gym Activities**, comes from this website: https://catalog.data.gov/dataset/recreation-spring-open-gym-activities. It tells us about different activities, where they happen when they occur, and who can join. Using this data, we could see what activities are available when they happen, and for whom they are meant, all during the spring season.

ABSTRACT:

The project revolves around analyzing data related to Recreational Spring Open Gym Activities, sourced from Kaggle. The goal is to gain insights into various activities, their locations, timings, and target participants during the spring season. The data is processed through Extract, Transform, Load (ETL) processes, normalized, and visualized to offer a comprehensive understanding of recreational activities.

PROBLEM STATEMENT:

The dataset initially presented challenges related to cleanliness, normalization, and visualization. Issues included combined address fields, redundant information, and the need for a structured data model. Addressing these challenges was crucial for deriving meaningful insights and presenting information effectively.

1. WHY IS THIS TABLE NOT IN 1NF?

Activity	Recreation Center	Address	Phone Numb	Age Require	Days of Week	Times	Address/Location
Volleyball	Praisner Community Center	14906 Old Columbia Pike	240-777-4970	18 & Up	Thursday	7:00pm - 9:30pm	14906 Old
Basketball	Upper County Community Center	8201 Emory Grove Rd	240-777-8077	17 & Under	Monday & Wednesday	6:00pm - 8:00pm	8201 Emory Grove
Basketball	North Potomac Community Center	13850 Travilah Rd	240-773-4800	All Ages	Sunday	12:00pm - 4:00pm	13850 Travilah Rd
Pickleball	Upper County Community Center	8201 Emory Grove Rd	240-777-8077	18 & Up	Monday	10:00am - 12:00pm	8201 Emory Grove
Badminton	North Potomac Community Center	13850 Travilah Rd	240-773-4800	17 & Under	Thursday	3:00pm - 6:00pm	13850 Travilah Rd

- In the given data, the house number and street address are combined in the same cell. This violates the principle of atomicity since a cell should only contain a single value.
- Two address columns suggest the presence of a repeating group, as each row should ideally represent a single entity. So, we have removed 1 address column to eliminate redundancy and conform to the first normal form (1NF). And, having two addresses indicates that there is a repeating group, as each should represent a single entity.

We have deleted the redundant address column and retained only a single column for the address, then we have removed the redundancy and taken a step toward conforming to the first normal form (1NF).

Activity	Recreation Center	Phone Numb	Age Require	Days of Week	Times	Address/Location
Volleyball	Praisner Community Center	240-777-4970	18 & Up	Thursday	7:00pm - 9:30pm	14906 Old
Basketball	Upper County Community Center	240-777-8077	17 & Under	Monday & Wednesday	6:00pm - 8:00pm	8201 Emory Grove
Basketball	North Potomac Community Center	240-773-4800	All Ages	Sunday	12:00pm - 4:00pm	13850 Travilah Rd
Pickleball	Upper County Community Center	240-777-8077	18 & Up	Monday	10:00am - 12:00pm	8201 Emory Grove
Badminton	North Potomac Community Center	240-773-4800	17 & Under	Thursday	3:00pm - 6:00pm	13850 Travilah Rd

2. 0NF – 1NF

Activity	Recreation Center	Phone Numb	Age Require	Days of Week	Times	Address/Location
Volleyball	Praisner Community Center	240-777-4970	18 & Up	Thursday	7:00pm - 9:30pm	14906 Old
Basketball	Upper County Community Center	240-777-8077	17 & Under	Monday & Wednesday	6:00pm - 8:00pm	8201 Emory Grove
Basketball	North Potomac Community Center	240-773-4800	All Ages	Sunday	12:00pm - 4:00pm	13850 Travilah Rd
Pickleball	Upper County Community Center	240-777-8077	18 & Up	Monday	10:00am - 12:00pm	8201 Emory Grove
Badminton	North Potomac Community Center	240-773-4800	17 & Under	Thursday	3:00pm - 6:00pm	13850 Travilah Rd

- We have restructured the data into separate columns for the start time, end time, house number, street, latitude, and longitude.

A	B	C	D	E	F	G	H	I	J	K	L	M
Activity	Recreation Center	House Number	Street Address	Latitude	Longitude	Phone Number	Age Requirements	Days of Week	Start Time	End Time	Recreation_info_id	Activitivity_schedule_id
A	B	C	D	E	F	G	H	I	J	K	L	M
Volleyball	Praisner Community Center	14906	Old Columbia Pike	-76.940764	39.101353	240-777-4970	18 & Up	Thursday	7:00pm	9:30pm	1	11
Basketball	Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	17 & Under	Monday	6:00pm	8:00pm	2	12
Basketball	Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	17 & Under	Wednesday	6:00pm	8:00pm	3	13
Basketball	North Potomac Community Center	13850	Travilah Rd	-77.232348	39.082916	240-773-4800	All Ages	Sunday	12:00pm	4:00pm	4	14
Pickleball	Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	18 & Up	Monday	10:00am	12:00pm	5	15
Badminton	North Potomac Community Center	13850	Travilah Rd	-77.232348	39.082916	240-773-4800	17 & Under	Thursday	3:00pm	6:00pm	6	16

FIND ALL FUNCTIONAL DEPENDENCY:

A	B	C	D	E	F	G	H	I	J	K	L	M
Activity	Recreation Center	House Number	Street Address	Latitude	Longitude	Phone Number	Age Requirements	Days of Week	Start Time	End Time	Recreation_info_id	Activitivity_schedule_id
A	B	C	D	E	F	G	H	I	J	K	L	M
Volleyball	Praisner Community Center	14906	Old Columbia Pike	-76.940764	39.101353	240-777-4970	18 & Up	Thursday	7:00pm	9:30pm	1	11
Basketball	Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	17 & Under	Monday	6:00pm	8:00pm	2	12
Basketball	Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	17 & Under	Wednesday	6:00pm	8:00pm	3	13
Basketball	North Potomac Community Center	13850	Travilah Rd	-77.232348	39.082916	240-773-4800	All Ages	Sunday	12:00pm	4:00pm	4	14
Pickleball	Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	18 & Up	Monday	10:00am	12:00pm	5	15
Badminton	North Potomac Community Center	13850	Travilah Rd	-77.232348	39.082916	240-773-4800	17 & Under	Thursday	3:00pm	6:00pm	6	16

B,L-> C,D,E,F,G

A,I,J,K,L,M ->H

Find the minimum cover:

- B,L-> C,D,E,F,G
- A,I,J,K,L,M->H

Step 1: Remove trivial FD.

None Found.

Step 2: Reduce the Right-Hand Side.

- B,L-> C
- B,L->D
- B,L->E
- B,L->F
- B,L ->G
- A,I,J,K,L,M -> H

Step 3: Reduce the Left-Hand Side.

- B-> C
- B->D
- B->E
- B->F
- B->G
- L-> C
- L->D
- L->E
- L->F
- L->G
- A->H
- I->H
- J->H
- K->H
- L->H

- M->H

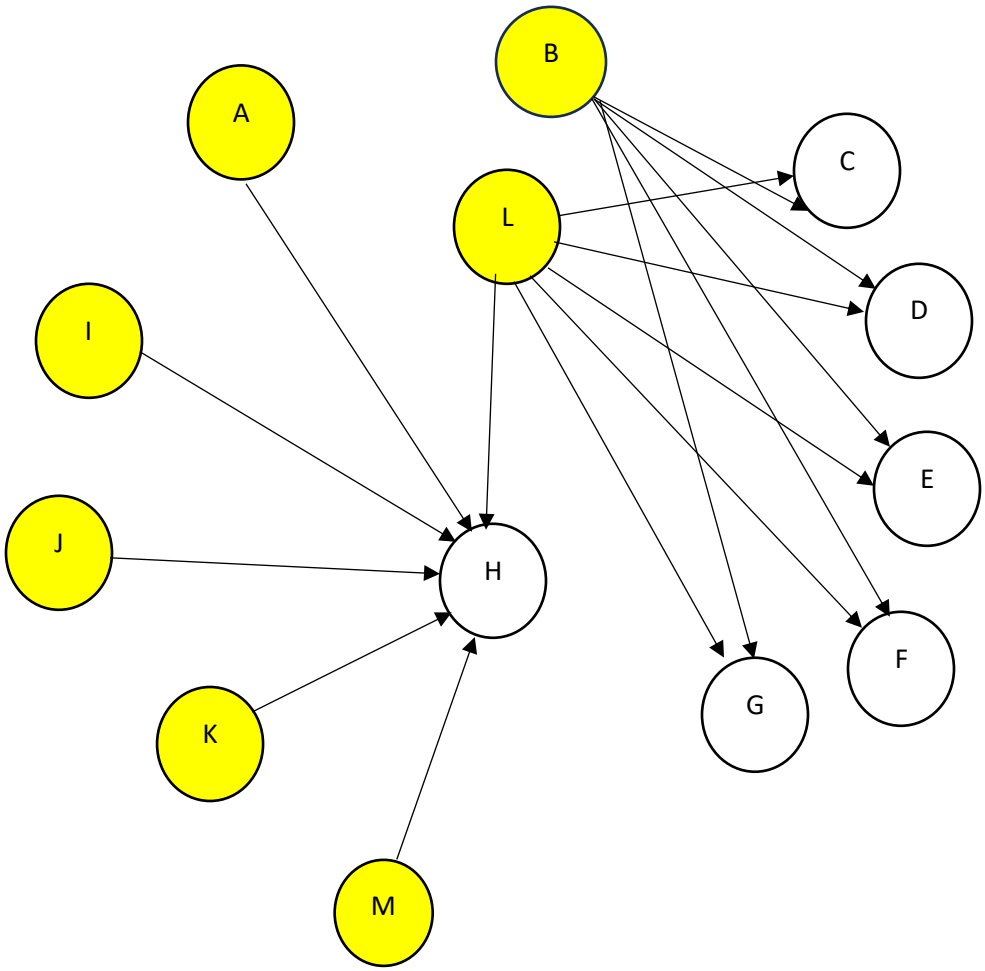
Step 4: Eliminate Redundancy.

- B-> C
- B->D
- B->E
- B->F
- B->G
- L-> C
- L->D
- L->E
- L->F

- L->G
- A->H
- I->H
- J->H
- K->H
- L->H
- M->H

SUPER KEY:

- B-> C
- B->D
- B->E
- B->F
- B->G
- L-> C
- L->D
- L->E
- L->F
- L->G
- A->H
- I->H
- J->H
- K->H
- L->H
- M->H



2 NF:

It is in 1NF.

Now, each relation is in 2NF because all non-prime attributes in each relation depend on the entire candidate key. There is no partial dependency, hence it is fully dependent on candidate keys.

R(ABDEFGHIJKLM) => R(<u>BL</u>CDEFG) + R(<u>AIJKLM</u>H)

Now, it's in 2NF.

3NF:

It's in 2NF.

There is no transitive dependency. Hence, it is in 3NF form.

FINAL RELATIONS:

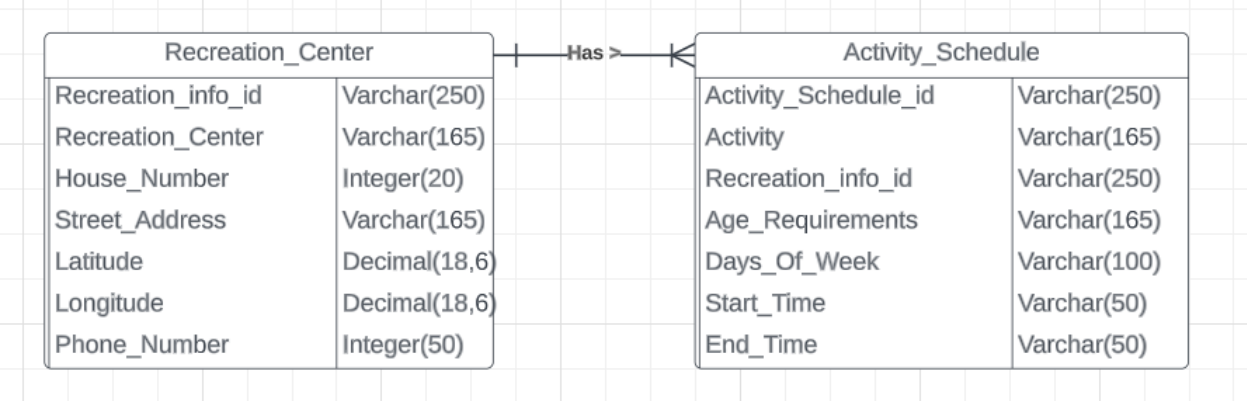
R(BLCDEFG)

Recreation Center	House Number	Street Address	Latitude	Longitude	Phone Number	Recreation_info_id
B	C	D	E	F	G	L
Praisner Community Center	14906	Old Columbia Pike	-76.940764	39.101353	240-777-4970	1
Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	2
Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	3
North Potomac Community Center	13850	Travilah Rd	-77.232348	39.082916	240-773-4800	4
Upper County Community Center	8201	Emory Grove Rd	-77.168899	39.149651	240-777-8077	5
North Potomac Community Center	13850	Travilah Rd	-77.232348	39.082916	240-773-4800	6

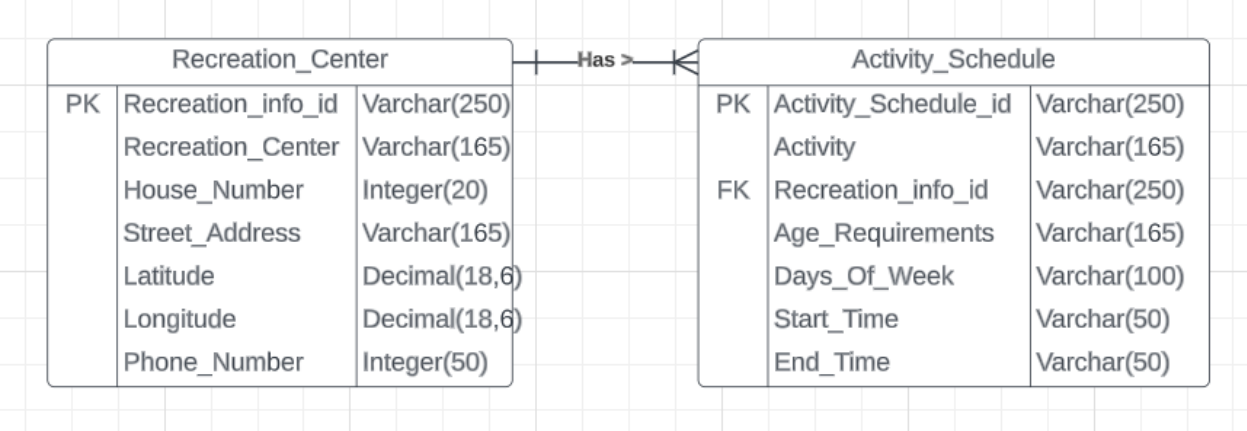
R(AIJKMLH)

Activity	Age Requirements	Days of Week	Start Time	End Time	Recreation_info_id	Activitivity_schedule_id
A	H	I	J	K	L	M
Volleyball	18 & Up	Thursday	7:00pm	9:30pm	1	11
Basketball	17 & Under	Monday	6:00pm	8:00pm	2	12
Basketball	17 & Under	Wednesday	6:00pm	8:00pm	3	13
Basketball	All Ages	Sunday	12:00pm	4:00pm	4	14
Pickleball	18 & Up	Monday	10:00am	12:00pm	5	15
Badminton	17 & Under	Thursday	3:00pm	6:00pm	6	16

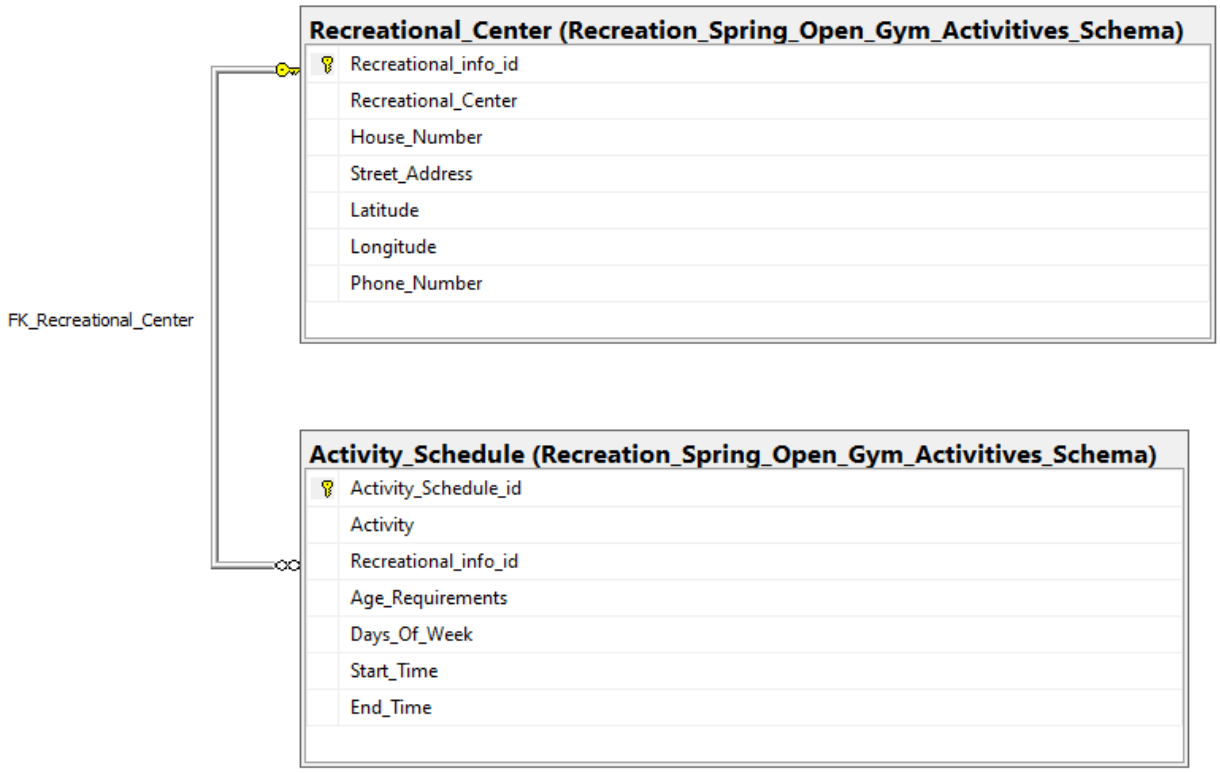
CONCEPTUAL DATA MODEL:



LOGICAL DATA MODEL:

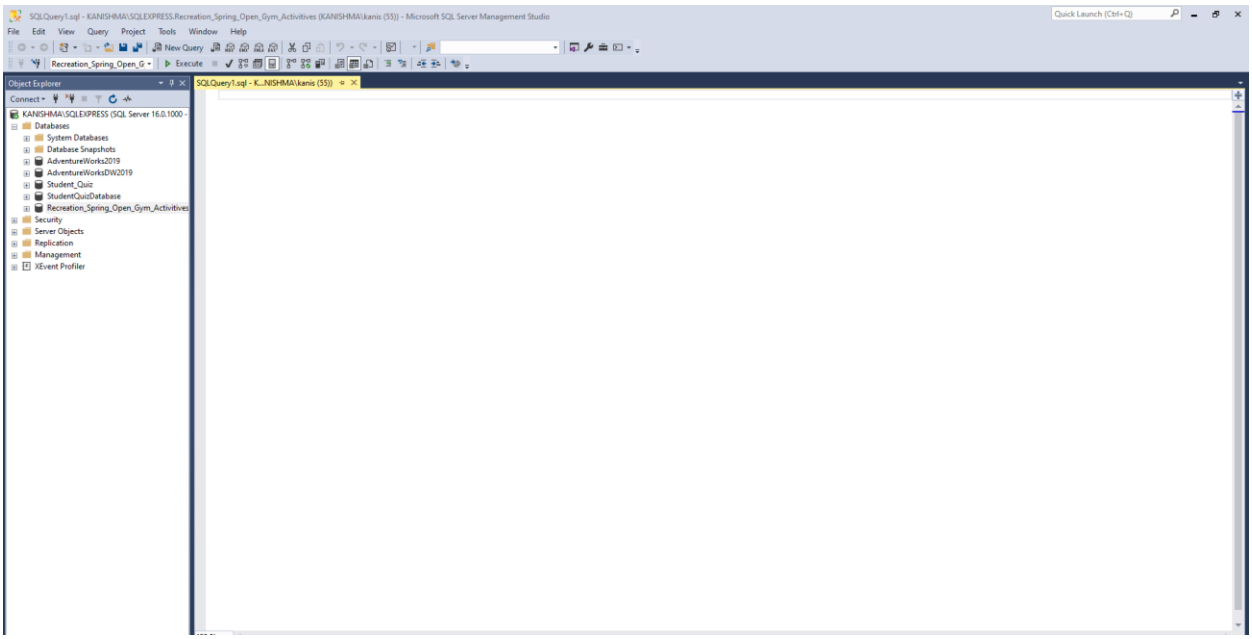


PHYSICAL DATA MODEL:

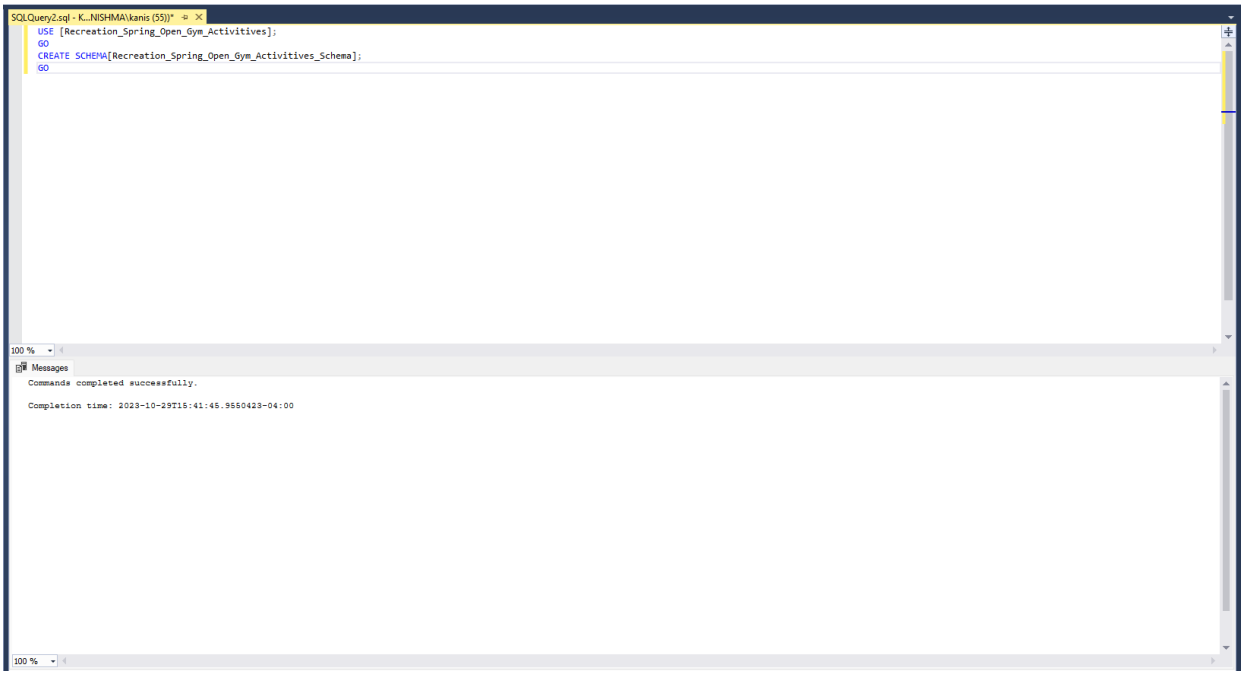


DDL Statement:

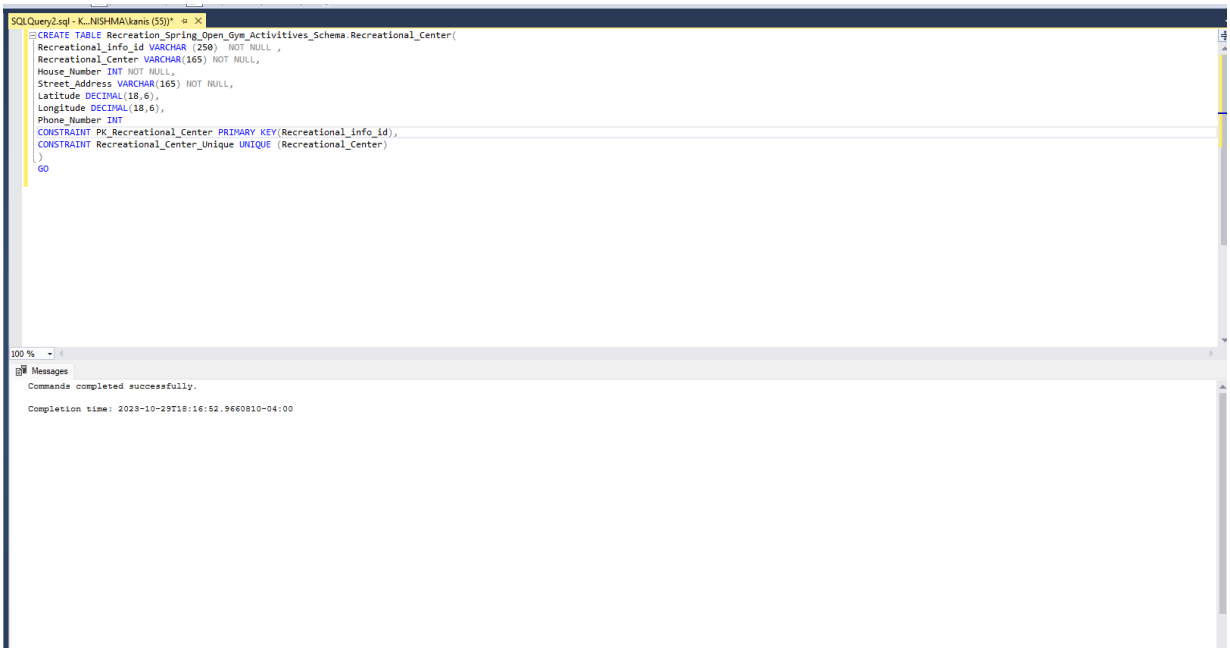
1) CREATE DATABASE:

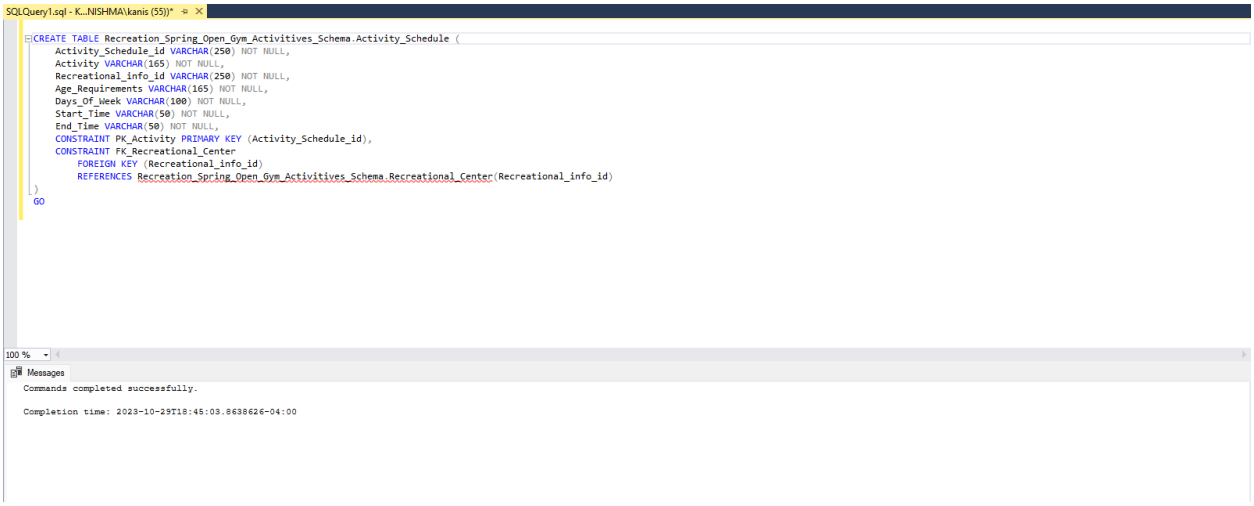


2) CREATE SCHEMA:



3) CREATE A DDL STATEMENT FOR ALL NORMALIZED TABLES, WITH PRIMARY AND FOREIGN KEYS:





Physical Tables in SQL Server Database:

Column Name	Data Type	Allow Nulls
Recreational_info_id	varchar(250)	☐
Recreational_Center	varchar(165)	☐
House_Number	int	☐
Street_Address	varchar(165)	☐
Latitude	decimal(18, 6)	☐
Longitude	decimal(18, 6)	☐
Phone_Number	int	☐
		☐

Column Name	Data Type	Allow Nulls
Activity_Schedule_id	varchar(250)	☐
Activity	varchar(165)	☐
Recreational_info_id	varchar(250)	☐
Age_Requirements	varchar(165)	☐
Days_Of_Week	varchar(100)	☐
Start_Time	varchar(50)	☐
End_Time	varchar(50)	☐
		☐

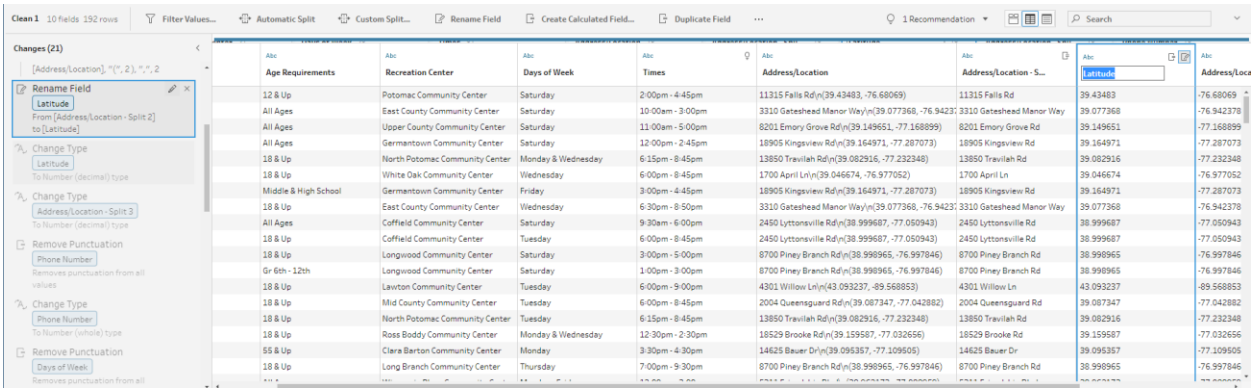
These are the 2 physical tables we created in the SQL Server Database.

Workflow of the ETL MODEL:

Our project has chosen to follow the ETL data modeling, and the workflow is as follows.

RENAME:

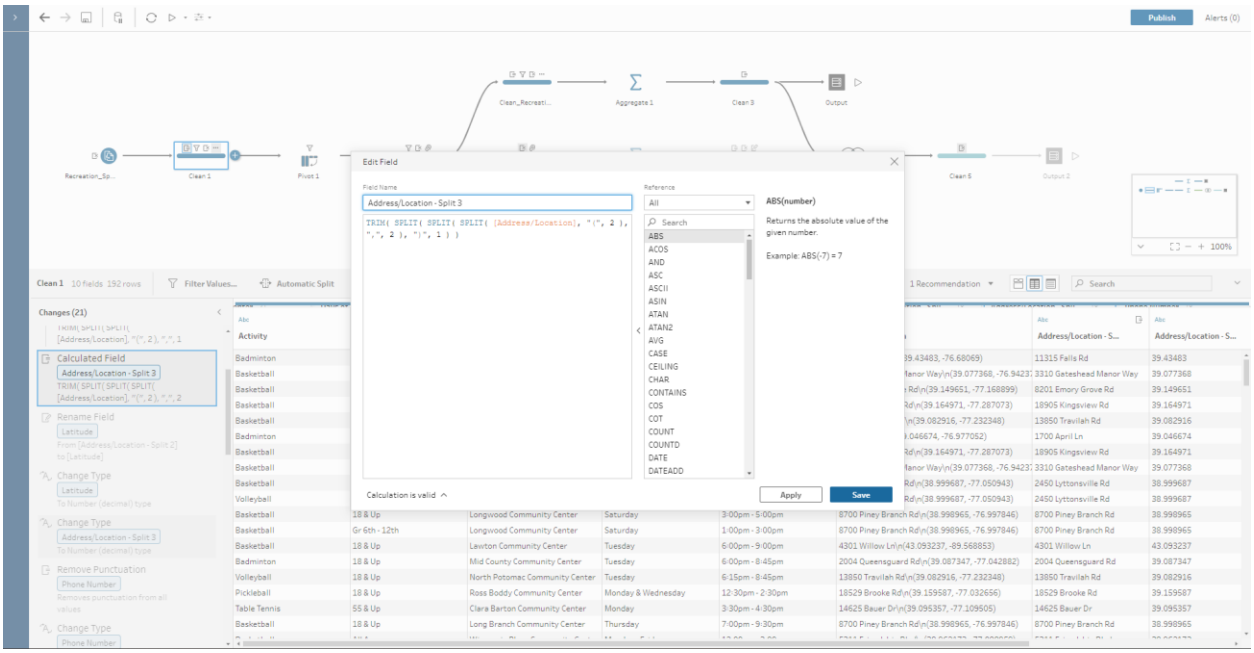
From the below screenshot, we have renamed the Address/Location – Split 2 field to Latitude field.



DATA CLEAN_UP:

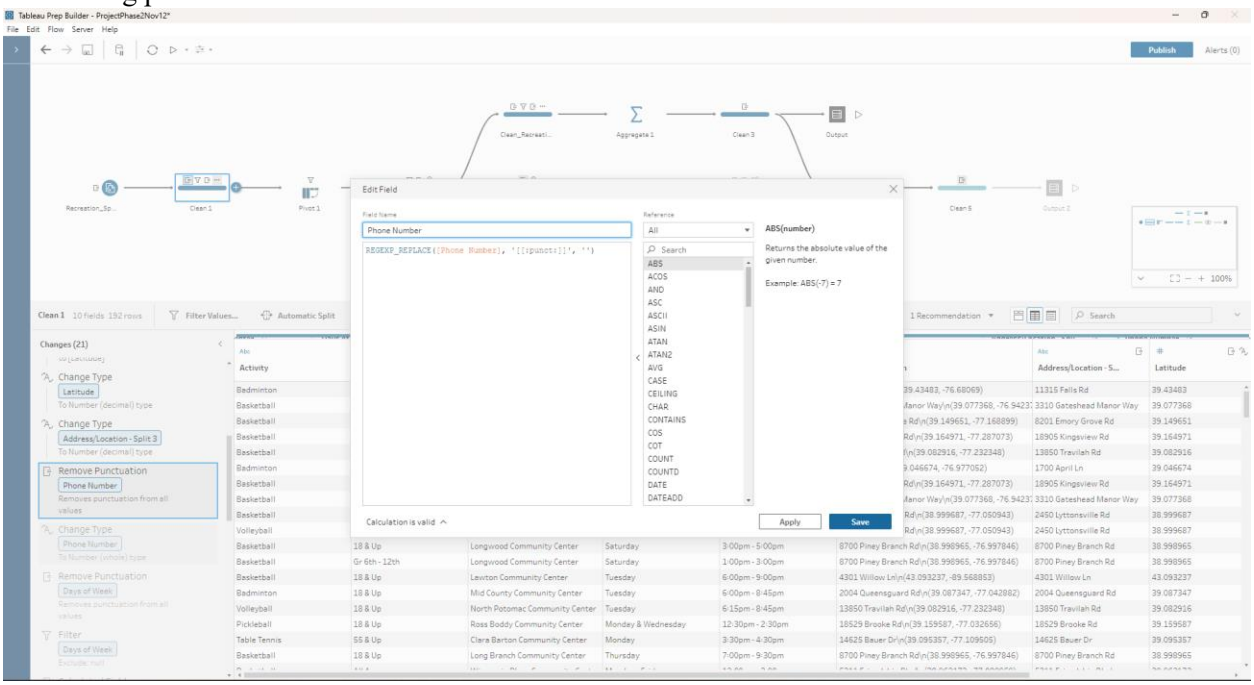
TRIM:

The below picture shows that we have cleaned up the trim spaces that has been unwantedly equipped by the Address/Location – Split 3 Attribute. Thus, Spaces have been removed from beginning and end of the field.



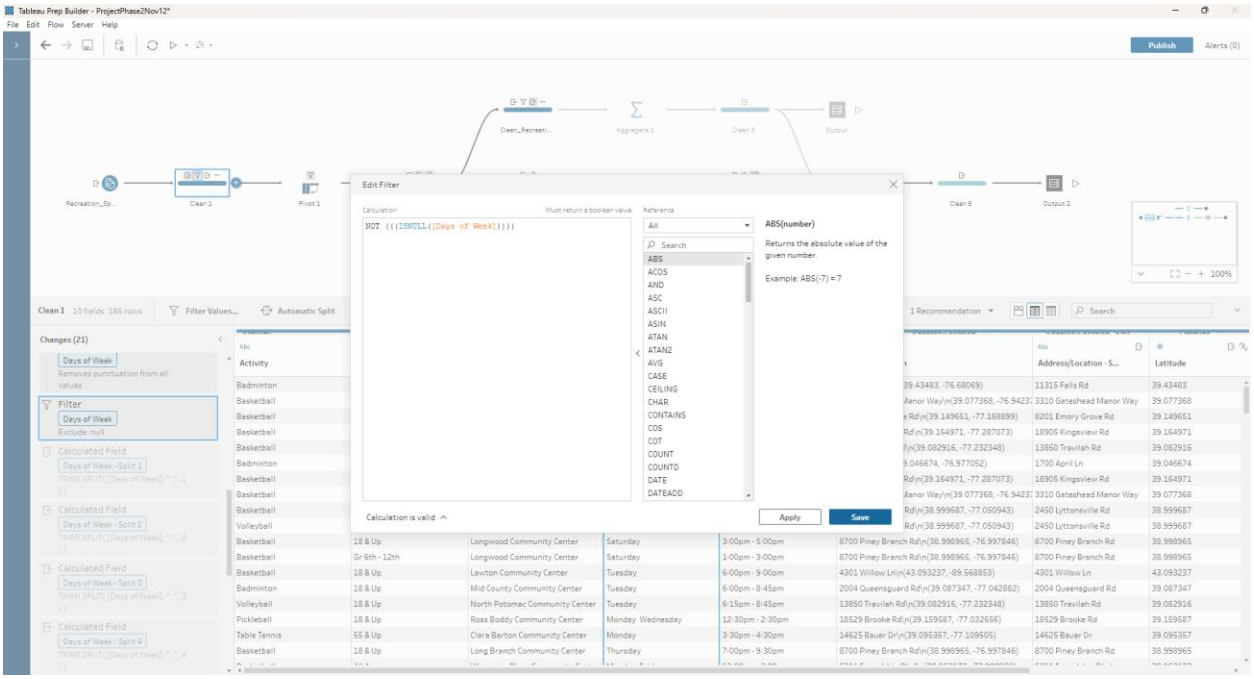
REMOVE PUNCTUATIONS:

Removing punctuation is also one of the clean up step. Where we could get the entire column by eliminating punctuations from it.



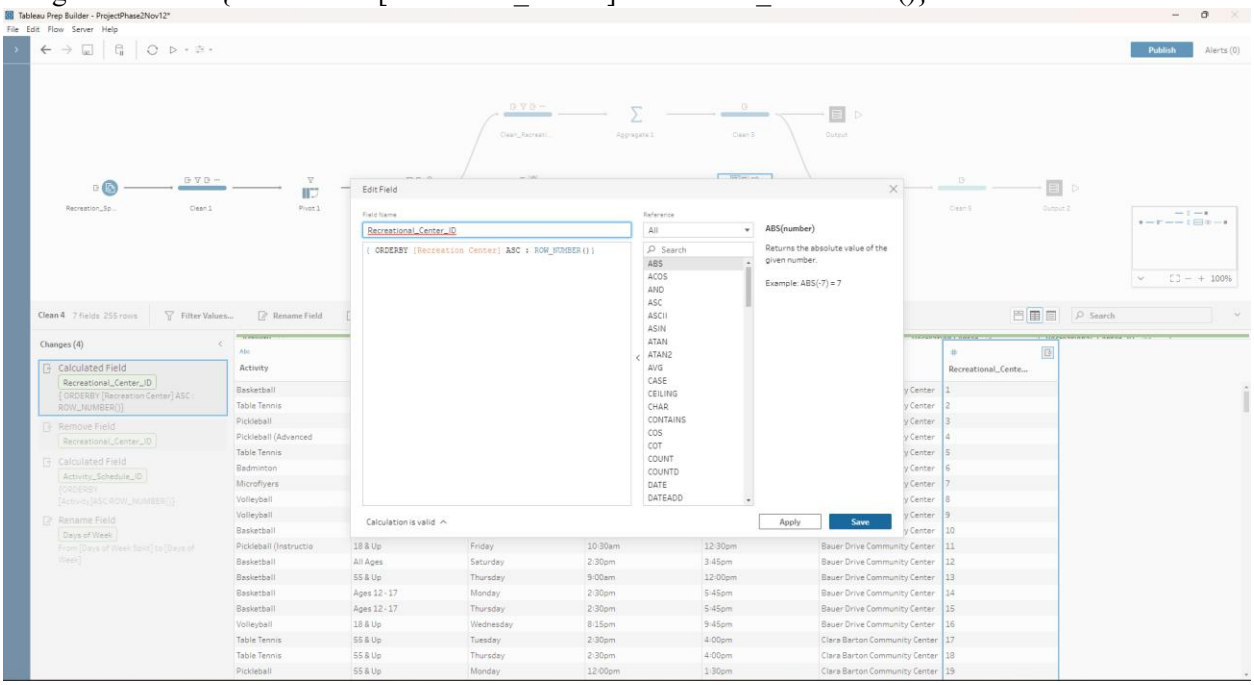
FILTERING DATA:

We need to filter out the null values from the list, in order to do that we have used Exclude by using filter option.



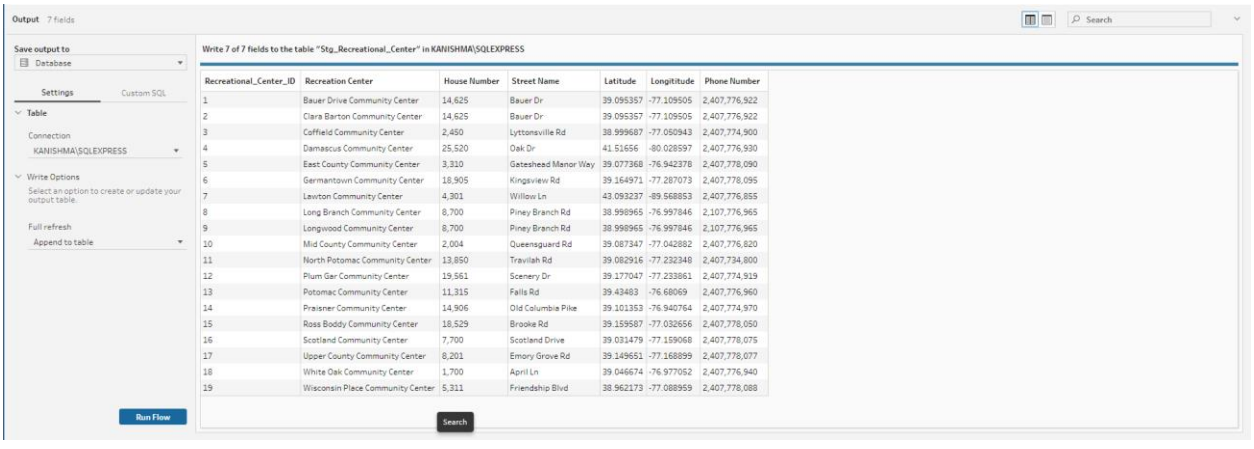
CALCULATED FIELD:

To add an auto generated field to construct Recreational_Center_ID, we have created a calculated field by using the formula {ORDERBY [Recreation_Center] ASC:ROW_NUMBER()}.

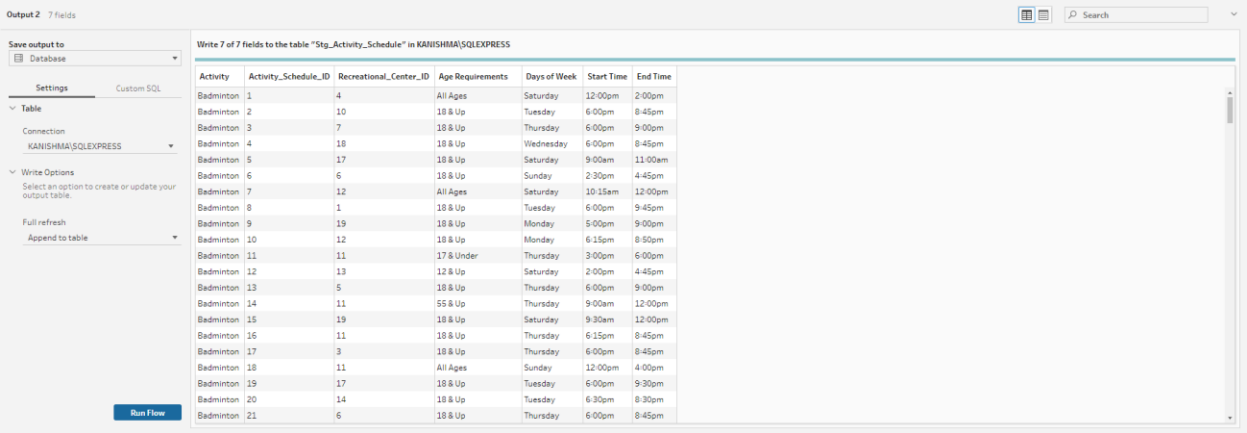


LOADING THE DATA TO THE TABLES:

At last, after cleaning, joining and aggregation process are dataset is ready to get deployed. We have connected to the MICROSOFT SQL SERVER, to load the data directly from staging tables to normalized tables. Below given are the screenshots.



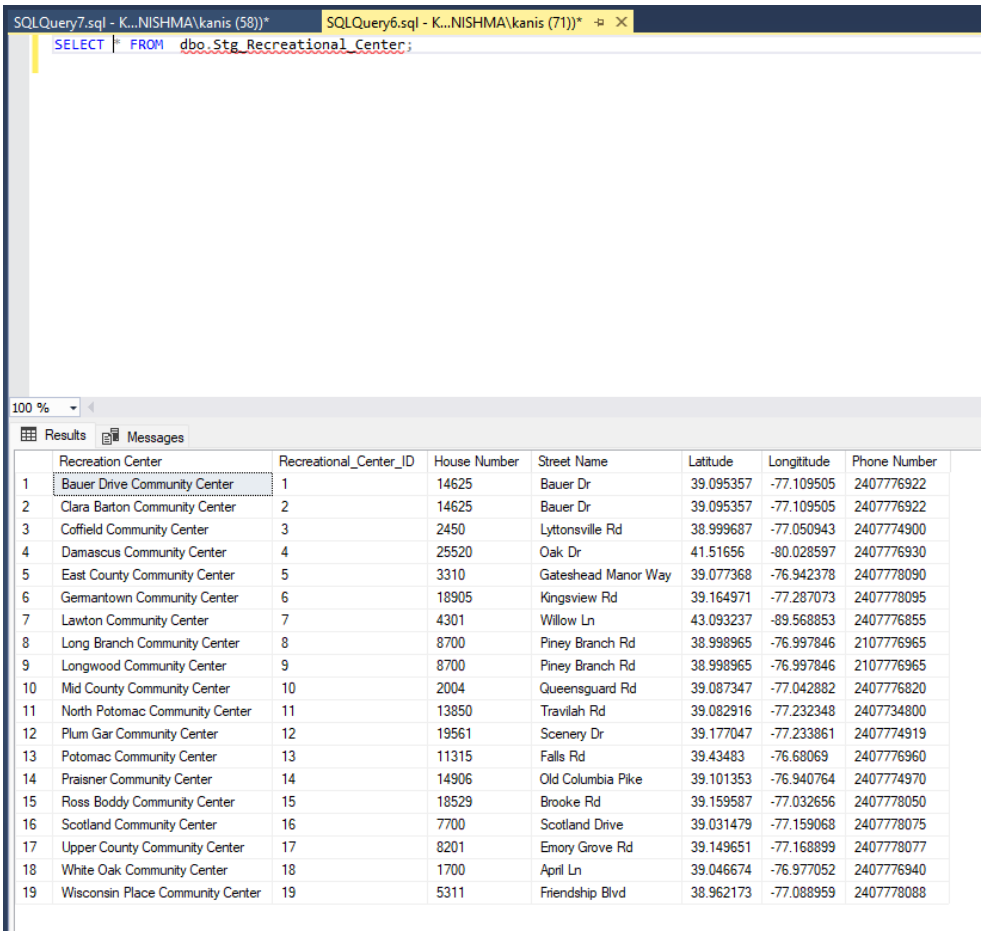
Recreational_Center_ID	Recreation Center	House Number	Street Name	Latitude	Longitude	Phone Number
1	Bauer Drive Community Center	14,625	Bauer Dr	39.095357	-77.109505	2,407,776,922
2	Clara Barton Community Center	14,625	Bauer Dr	39.095357	-77.109505	2,407,776,922
3	Coffield Community Center	2,450	Lyttonsville Rd	38.999687	-77.050943	2,407,774,900
4	Damascus Community Center	25,520	Oak Dr	41.51656	-80.028597	2,407,776,930
5	East County Community Center	3,310	Gateshead Manor Way	39.077368	-76.942378	2,407,778,090
6	Germantown Community Center	18,905	Kingsview Rd	39.164971	-77.287073	2,407,778,095
7	Lawton Community Center	4,301	Willow Ln	43.093237	-89.568853	2,407,776,855
8	Long Branch Community Center	8,700	Piney Branch Rd	38.998965	-76.997846	2,107,776,965
9	Longwood Community Center	8,700	Piney Branch Rd	38.998965	-76.997846	2,107,776,965
10	Mid County Community Center	2,004	Queensguard Rd	39.087347	-77.042882	2,407,776,820
11	North Potomac Community Center	13,850	Travilah Rd	39.082916	-77.232348	2,407,734,800
12	Plum Gar Community Center	19,561	Scenery Dr	39.177047	-77.233861	2,407,774,919
13	Potomac Community Center	11,315	Falls Rd	39.43483	-76.68069	2,407,776,960
14	Praisner Community Center	14,906	Old Columbia Pike	39.101353	-76.940764	2,407,774,970
15	Ross Boddy Community Center	18,529	Brooke Rd	39.159587	-77.032656	2,407,778,050
16	Scotland Community Center	7,700	Scotland Drive	39.031479	-77.159068	2,407,778,075
17	Upper County Community Center	8,201	Emory Grove Rd	39.149651	-77.168899	2,407,778,077
18	White Oak Community Center	1,700	April Ln	39.046674	-76.977052	2,407,776,940
19	Wisconsin Place Community Center	5,311	Friendship Blvd	38.962173	-77.088959	2,407,778,088



RESULTS OF LOADING in SQL Server:

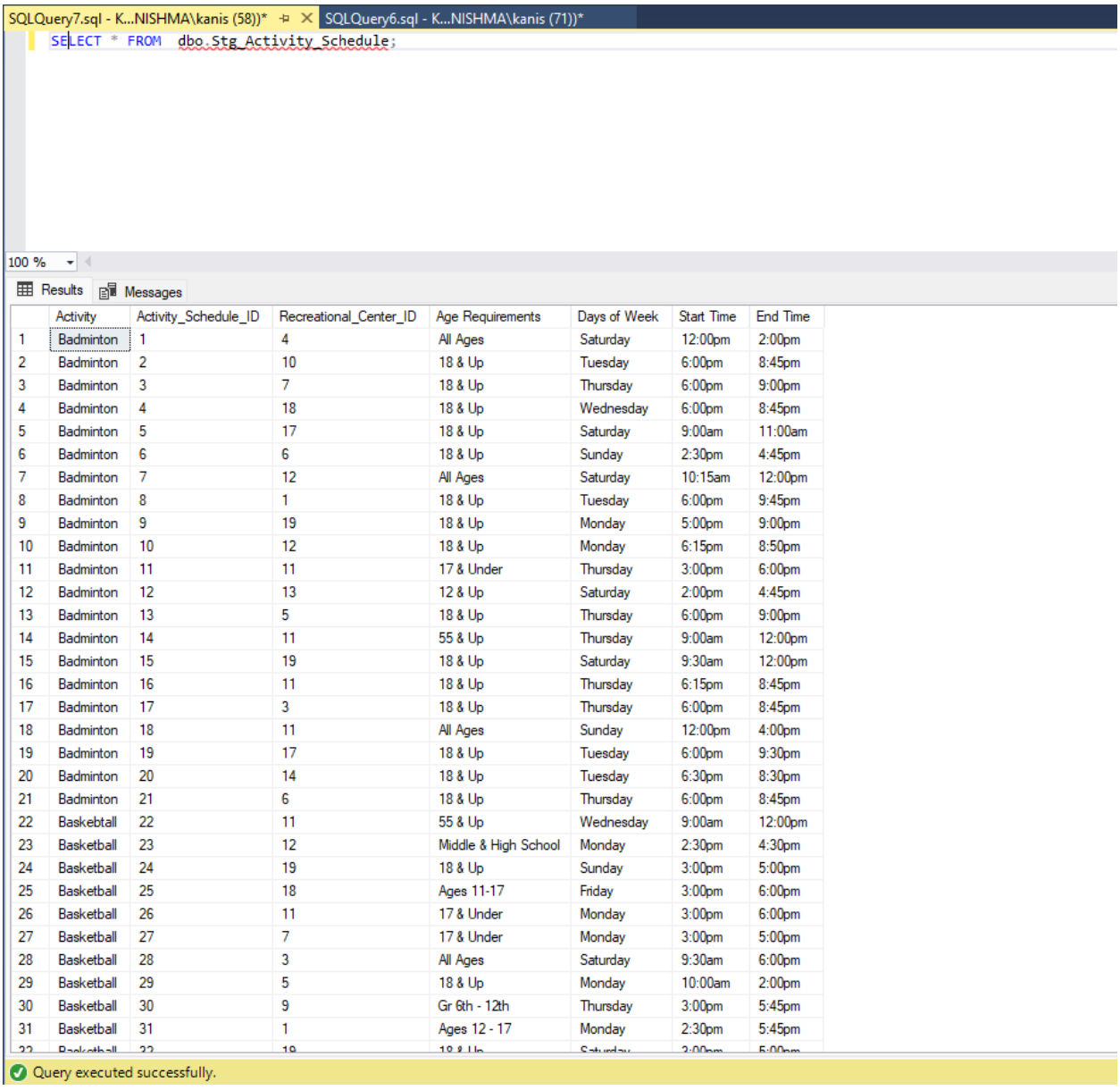
By using a select query, the results show that the raw data has been cleaned and loaded into the finalized two tables and they are:

RECREATION TABLE:



	Recreation Center	Recreational_Center_ID	House Number	Street Name	Latitude	Longitude	Phone Number
1	Bauer Drive Community Center	1	14625	Bauer Dr	39.095357	-77.109505	2407776922
2	Clara Barton Community Center	2	14625	Bauer Dr	39.095357	-77.109505	2407776922
3	Coffield Community Center	3	2450	Lyttonsville Rd	38.999687	-77.050943	2407774900
4	Damascus Community Center	4	25520	Oak Dr	41.51656	-80.028597	2407776930
5	East County Community Center	5	3310	Gateshead Manor Way	39.077368	-76.942378	2407778090
6	Germantown Community Center	6	18905	Kingsview Rd	39.164971	-77.287073	2407778095
7	Lawton Community Center	7	4301	Willow Ln	43.093237	-89.568853	2407776855
8	Long Branch Community Center	8	8700	Piney Branch Rd	38.998965	-76.997846	2107776965
9	Longwood Community Center	9	8700	Piney Branch Rd	38.998965	-76.997846	2107776965
10	Mid County Community Center	10	2004	Queensguard Rd	39.087347	-77.042882	2407776820
11	North Potomac Community Center	11	13850	Travilah Rd	39.082916	-77.232348	2407734800
12	Plum Gar Community Center	12	19561	Scenery Dr	39.177047	-77.233861	2407774919
13	Potomac Community Center	13	11315	Falls Rd	39.43483	-76.68069	2407776960
14	Praisner Community Center	14	14906	Old Columbia Pike	39.101353	-76.940764	2407774970
15	Ross Boddy Community Center	15	18529	Brooke Rd	39.159587	-77.032656	2407778050
16	Scotland Community Center	16	7700	Scotland Drive	39.031479	-77.159068	2407778075
17	Upper County Community Center	17	8201	Emory Grove Rd	39.149651	-77.168899	2407778077
18	White Oak Community Center	18	1700	April Ln	39.046674	-76.977052	2407776940
19	Wisconsin Place Community Center	19	5311	Friendship Blvd	38.962173	-77.088959	2407778088

Activity_ Schedule Table:



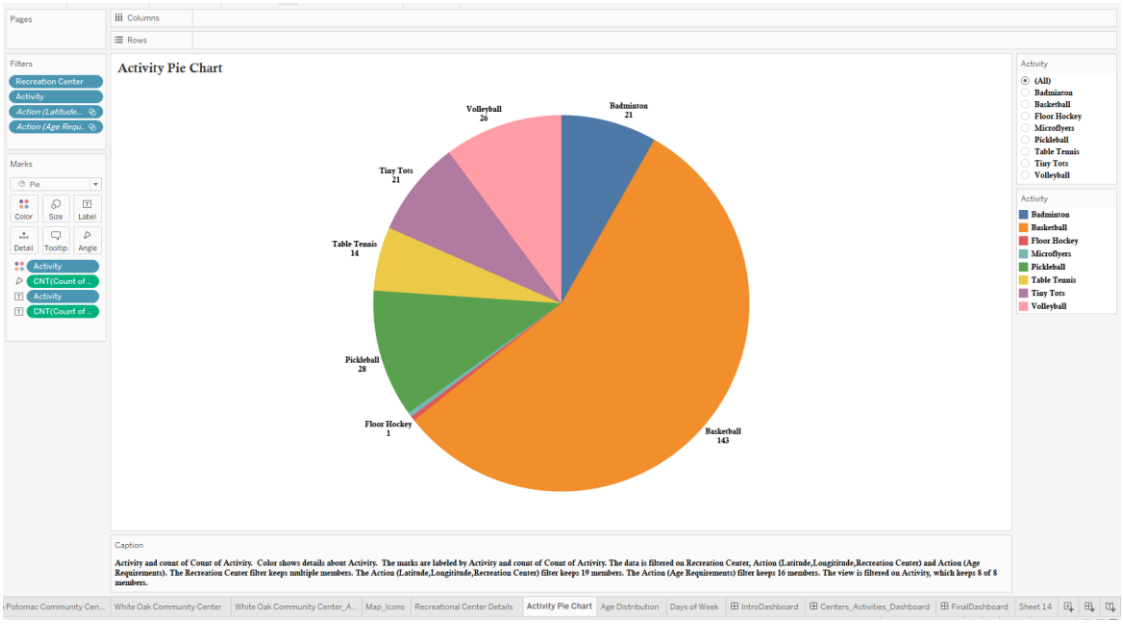
	Activity	Activity_Schedule_ID	Recreational_Center_ID	Age Requirements	Days of Week	Start Time	End Time
1	Badminton	1	4	All Ages	Saturday	12:00pm	2:00pm
2	Badminton	2	10	18 & Up	Tuesday	6:00pm	8:45pm
3	Badminton	3	7	18 & Up	Thursday	6:00pm	9:00pm
4	Badminton	4	18	18 & Up	Wednesday	6:00pm	8:45pm
5	Badminton	5	17	18 & Up	Saturday	9:00am	11:00am
6	Badminton	6	6	18 & Up	Sunday	2:30pm	4:45pm
7	Badminton	7	12	All Ages	Saturday	10:15am	12:00pm
8	Badminton	8	1	18 & Up	Tuesday	6:00pm	9:45pm
9	Badminton	9	19	18 & Up	Monday	5:00pm	9:00pm
10	Badminton	10	12	18 & Up	Monday	6:15pm	8:50pm
11	Badminton	11	11	17 & Under	Thursday	3:00pm	6:00pm
12	Badminton	12	13	12 & Up	Saturday	2:00pm	4:45pm
13	Badminton	13	5	18 & Up	Thursday	6:00pm	9:00pm
14	Badminton	14	11	55 & Up	Thursday	9:00am	12:00pm
15	Badminton	15	19	18 & Up	Saturday	9:30am	12:00pm
16	Badminton	16	11	18 & Up	Thursday	6:15pm	8:45pm
17	Badminton	17	3	18 & Up	Thursday	6:00pm	8:45pm
18	Badminton	18	11	All Ages	Sunday	12:00pm	4:00pm
19	Badminton	19	17	18 & Up	Tuesday	6:00pm	9:30pm
20	Badminton	20	14	18 & Up	Tuesday	6:30pm	8:30pm
21	Badminton	21	6	18 & Up	Thursday	6:00pm	8:45pm
22	Baskebtall	22	11	55 & Up	Wednesday	9:00am	12:00pm
23	Basketball	23	12	Middle & High School	Monday	2:30pm	4:30pm
24	Basketball	24	19	18 & Up	Sunday	3:00pm	5:00pm
25	Basketball	25	18	Ages 11-17	Friday	3:00pm	6:00pm
26	Basketball	26	11	17 & Under	Monday	3:00pm	6:00pm
27	Basketball	27	7	17 & Under	Monday	3:00pm	5:00pm
28	Basketball	28	3	All Ages	Saturday	9:30am	6:00pm
29	Basketball	29	5	18 & Up	Monday	10:00am	2:00pm
30	Basketball	30	9	Gr 6th - 12th	Thursday	3:00pm	5:45pm
31	Basketball	31	1	Ages 12 - 17	Monday	2:30pm	5:45pm

DATA VISUALIZATION:

Data visualization plays a crucial role in conveying insights from the processed data in a comprehensible and visually appealing manner. In this project, various visualization techniques were applied:

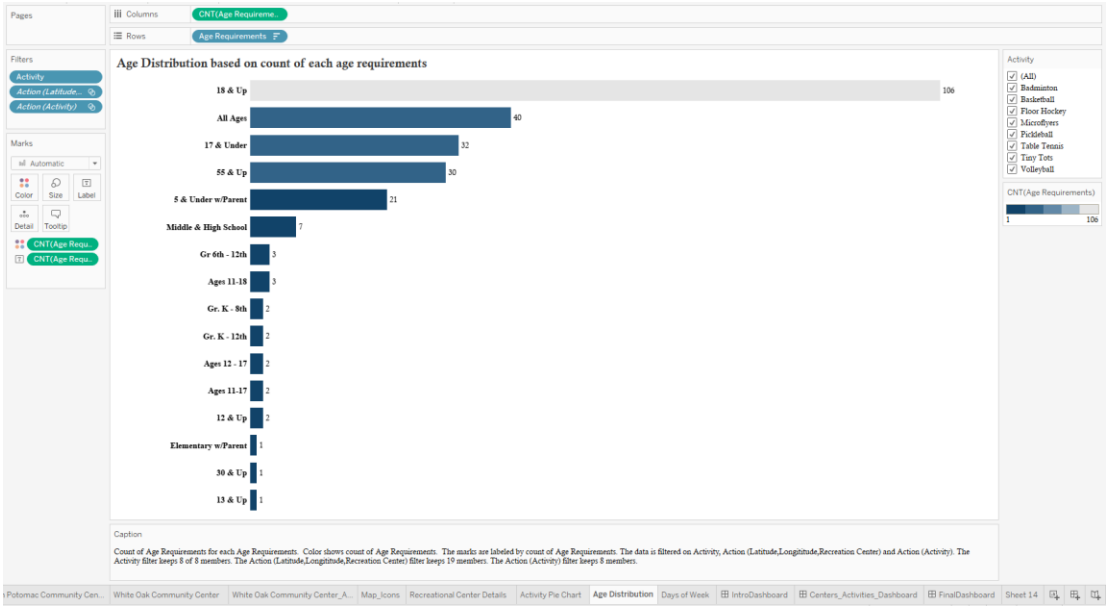
1. PIE CHART:

A pie chart was used to visually represent the frequency of each activity conducted in recreational centers. This chart offers a quick and intuitive understanding of the distribution of activities.



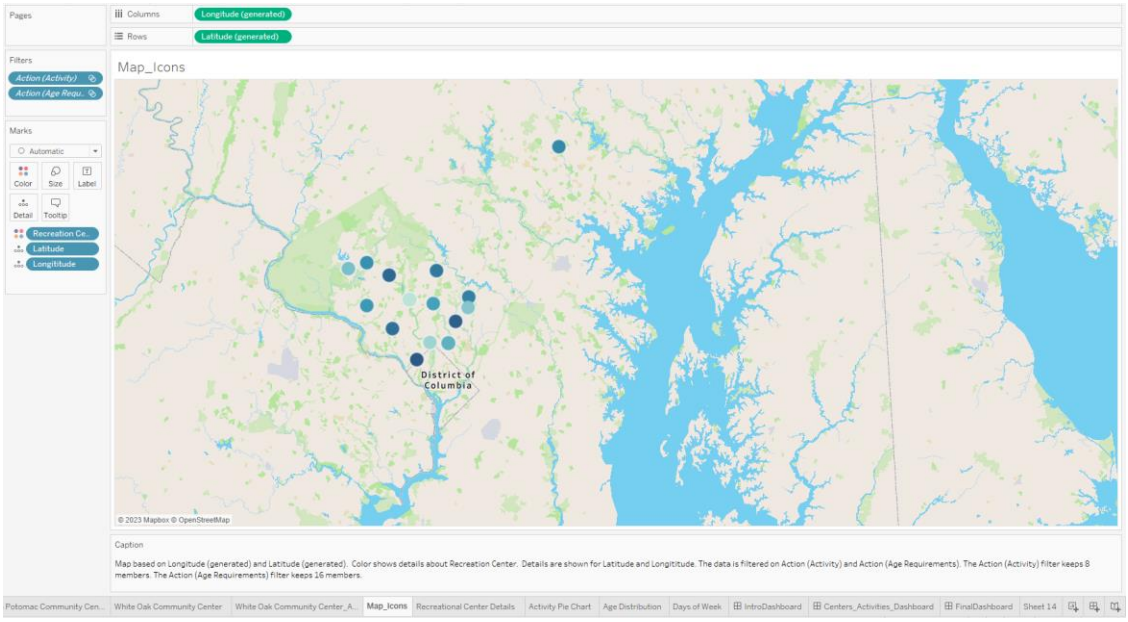
2. BAR CHART:

A bar chart was employed to display the age distribution based on the count of each age group. This visualization helps in understanding which age groups participate more frequently in recreational activities.



3. GEOGRAPHICAL MAP:

A geographical map was utilized to illustrate the areas with the highest concentration of recreational centers conducting activities. Each center's address details were incorporated using tooltips for additional information.

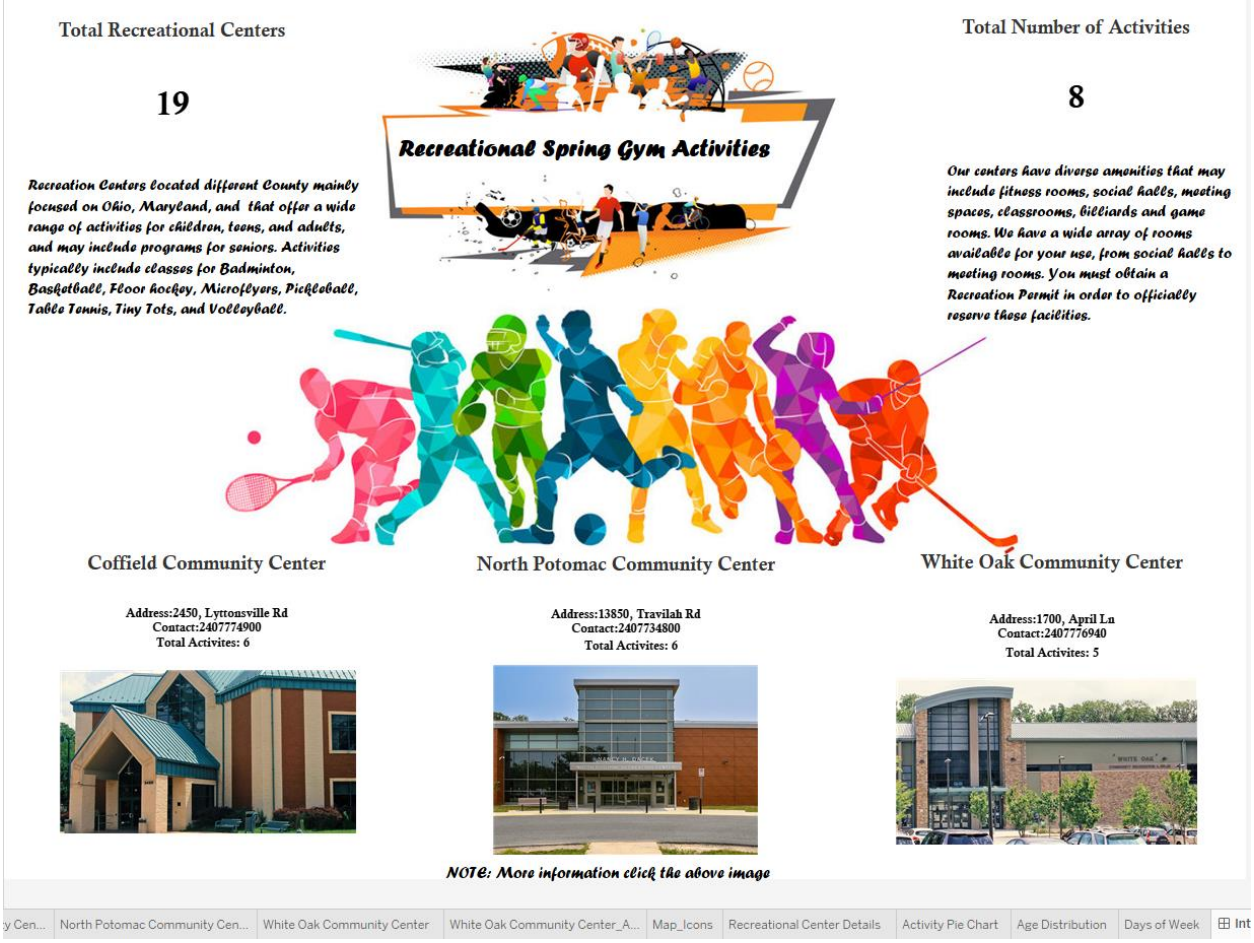


4. TABLE VISUALIZATION:

A highlighted table was employed to present the days of the week with the highest activity counts. This tabular format makes it easy to identify the peak days for recreational activities.

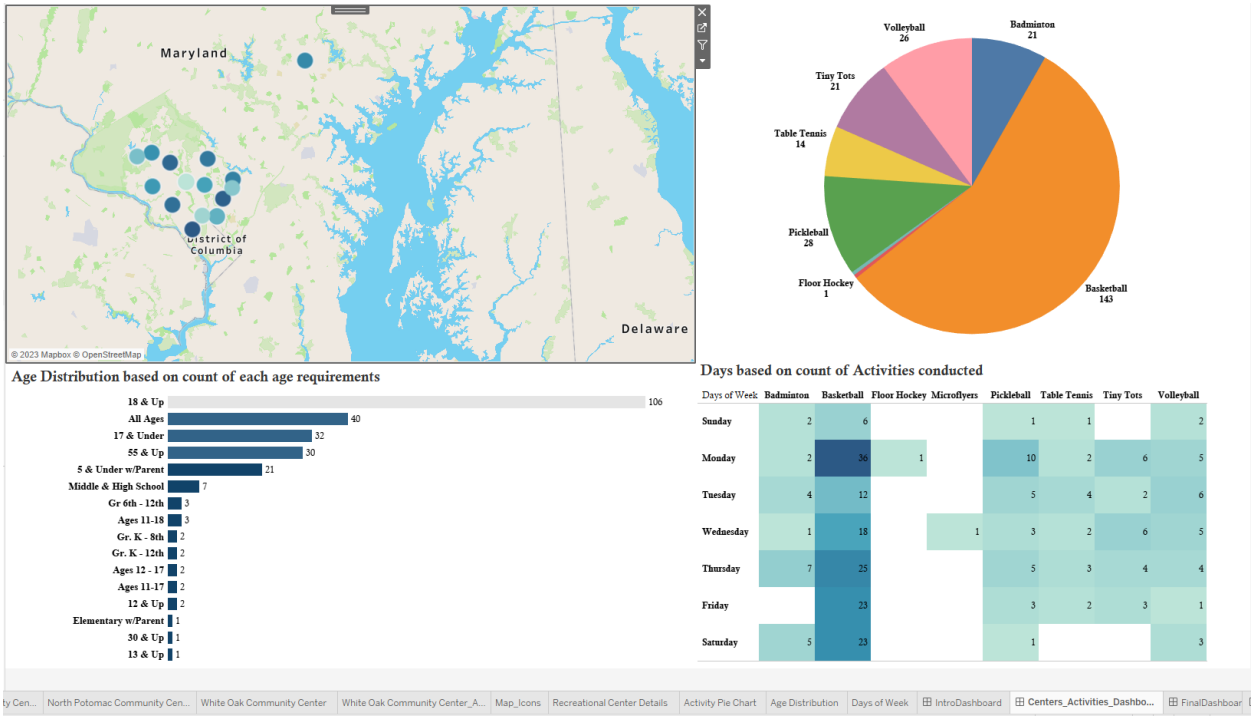
DASHBOARD: Introduction Dashboard

The recreational spring gym activities poster provides valuable insights into a variety of exercises and fun options, including badminton, basketball, floor hockey, micro flyers, pickleball, table tennis, tiny tots, and volleyball. It highlights the availability of these activities across different community centers in Pennsylvania, Wisconsin, and Maryland. The poster features 19 recreational centers offering these activities and mentions additional amenities like fitness rooms, social halls, meeting spaces, classrooms, and game rooms. Specific details about centers such as Coffield Community Center, North Potomac Community Center, and White Oak Community Center are provided, encouraging further exploration for information on activity schedules, costs, and any specific requirements. Overall, the poster emphasizes the positive aspects of engaging in recreational spring gym activities for exercise, enjoyment, and community involvement.



DASHBOARD: Recreational and Activity Dashboard

The Recreational and Activity Dashboard is a dynamic visualization tool providing a comprehensive overview of Recreational Spring Open Gym Activities. The geographical overview utilizes an interactive map and bar chart to showcase the distribution of recreational centers across states. The Recreation Center Details section employs a pie chart to highlight the frequency of activities in each center. Additionally, a bar chart illustrates age distribution based on specific age groups, and a table displays the days of the week with corresponding activity counts. The dashboard enhances understanding of activity popularity, geographic concentrations, and age preferences, fostering insightful analysis for diverse stakeholders involved in recreational planning.



Clicking on specific elements, such as a recreational center or age group, triggers a detailed information display. This interactive feature enhances user engagement and allows for a deeper exploration of the data. These visualization techniques collectively provide a comprehensive overview of recreational activities, their popularity, and key factors influencing their scheduling. The visuals serve as powerful tools for decision-makers to derive meaningful insights and make informed choices based on the data.

KEY FINDINGS AND INSIGHTS FROM THE DATASET:

- People living in the community can use visualization to find out which recreational activities are available when they happen, and if they match their interests or their children's interests.
- By looking at the chart, they can quickly see the types of activities, the busiest days, and which age groups are catered to, helping them decide where to participate.
- Recreation centers can utilize insights to optimize their program offerings, focusing on popular activities and peak days.
- They can identify specific age groups that may need more attention or tailored programs, enhancing engagement. Develop attractive programs for ages 13 & and up to increase participation.
- By understanding the geographical distribution of recreational centers, they can allocate resources and support where it is most needed.
- Introduce and promote new activities.
- Collaborate with centers, especially in regions with lower participation.

CONCLUSION:

In summary, the recreational spring gym activities dataset and visuals offer insights into the diverse exercise options at various community centers. The data highlights the variety of activities and amenities available, promoting community engagement. The user-friendly dashboards encourage exploration of specific centers and activities, contributing to a healthier and more connected community through accessible recreation.