

## **SQL FOR DATA SCIENCE FOR CAPSTONE PROJECT**

### **(MILESTONE 1)**

**Name: Okonkwo Maureen**

**Client: Sports Stats**

**Data set: Olympics Dataset**

**Date: April 2022**

#### **Question 1:**

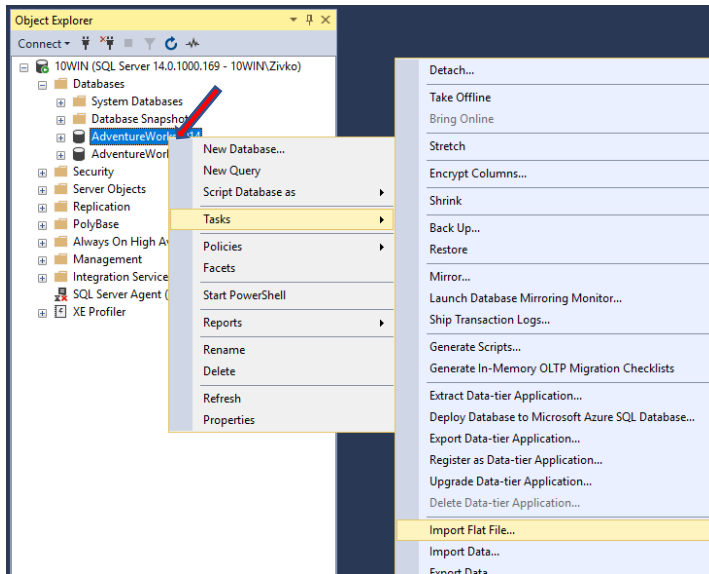
**The client and dataset respectfully are;** Sport stats and Olympics dataset and why I choose the client and dataset is because the dataset contains about 120 years of Olympics data and I am curious to know what is in it for me.

Steps on how to import data:

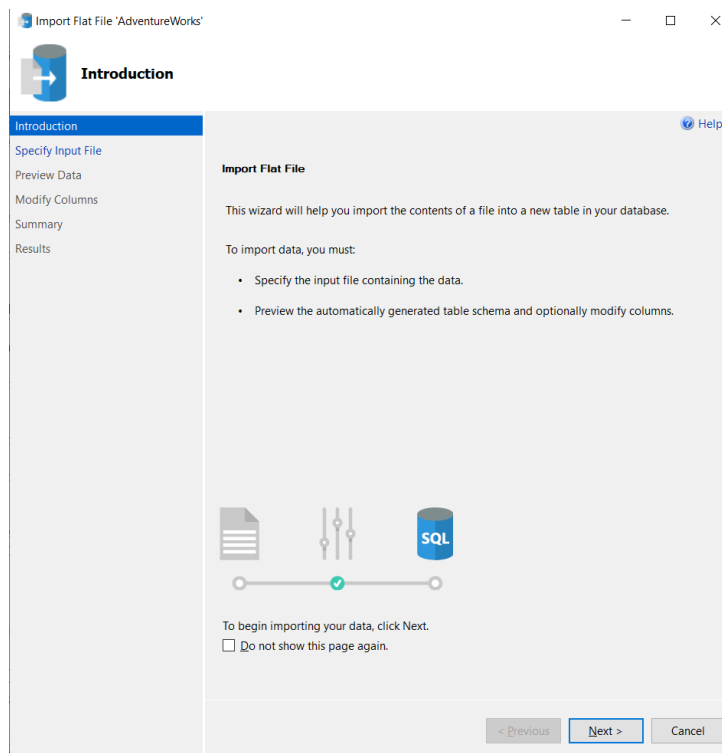
- 1)Downloaded the dataset from Coursera
- 2)Converted the zip file to a csv file (excel)
- 3)Import the csv file into my SQL Server in order to query the dataset

The steps are:

- a) Navigate to Object Explorer and select the data base you want to use, then import your flat file.



b) Click Next to proceed with the importing the flat file.



c) Set up the location of a flat file which will be used for importing into a SQL Server database and also add the table name and the schema.

**Import Flat File 'AdventureWorks'**

**Specify Input File**

Introduction  
**Specify Input File**  
Preview Data  
Modify Columns  
Summary  
Results

**Specify Input File**  
This operation will create a table from your input file.

Location of file to be imported  
C:\Users\maureen\OneDrive\Documents\COURSE\athlete\_event.csv **Browse...**

New table name:  
athlete\_event

Table schema:  
dbo

< Previous Next > Cancel

## D) Preview Data

**Import Flat File 'AdventureWorks'**

**Preview Data**

Introduction  
Specify Input File  
**Preview Data**  
Modify Columns  
Summary  
Results

**Preview Data**  
This operation analyzed the input file structure to generate the preview below for up to the first 50 rows.

ID	Name	Sex	Age	Height	Weight	Team
1	A. Djang	M	24	180	80	Chi
2	A. Lamusi	M	23	170	60	Chi
3	Gunn Nielse...	M	24			Denr
4	Edgar Linde...	M	34			Denr
5	Christine Jac...	F	21	185	82	Neeth
5	Christine Jac...	F	21	185	82	Neeth
5	Christine Jac...	F	25	185	82	Neeth
5	Christine Jac...	F	25	185	82	Neeth
5	Christine Jac...	F	27	185	82	Neeth
6	Per Knut Aal...	M	31	188	75	Unitr
6	Per Knut Aal...	M	31	188	75	Unitr
6	Per Knut Aal...	M	31	188	75	Unitr
6	Per Knut Aal...	M	33	188	75	Unitr
6	Per Knut Aal...	M	33	188	75	Unitr
6	Per Knut Aal...	M	33	188	75	Unitr
6	Per Knut Aal...	M	33	188	75	Unitr
7	John Aalberg	M	31	183	72	Unitr
7	John Aalberg	M	31	183	72	Unitr

☒ Use Rich Data Type Detection - may provide a closer type fit. However, cells with anomalous values may be dropped.

< Previous Next > Cancel

## E) Modify the data type, Primary Key and Allowing Null Values

Import Flat File 'AdventureWorks'

**Modify Columns**

Introduction  
Specify Input File  
Preview Data  
**Modify Columns**  
Summary  
Results

**Modify Columns**

This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.

Column Name	Data Type	Primary Key	Allow Nulls
ID	tinyint	<input type="checkbox"/>	<input type="checkbox"/>
Name	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Sex	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Age	tinyint	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Height	tinyint	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Weight	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Team	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
NOC	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Games	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Year	smallint	<input type="checkbox"/>	<input type="checkbox"/>
Season	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
City	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Sport	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Event	nvarchar(100)	<input type="checkbox"/>	<input type="checkbox"/>
Medal	nvarchar(50)	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Row granularity of error reporting (performance impact with smaller ranges) No Range

< Previous Next > Cancel

## G) The summary pages

Import Flat File 'AdventureWorks'

**Summary**

Introduction  
Specify Input File  
Preview Data  
Modify Columns  
**Summary**  
Results

**Summary**

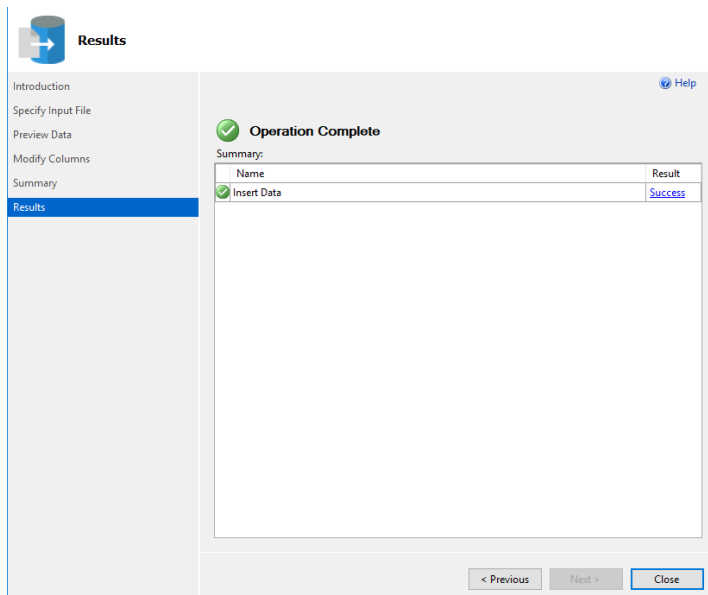
To complete the operation using the specified inputs, click Finish.

Import Information

- Name: LAPTOP-N0HMLPGA\SQLEXPRESS
- Database Name: AdventureWorks
- Table Name: dbo.athlete\_event
- File to be imported: C:\Users\maureen\OneDrive\Documents\COURSERA\athlete\_event.csv

< Previous Finish Cancel

## F) The result Page

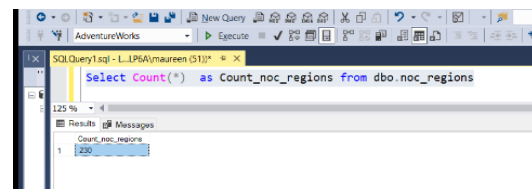
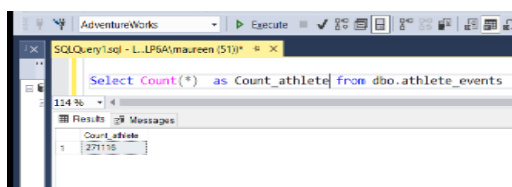


## CLEANING DATA:

Removing all duplicate tables and analyzing the table.

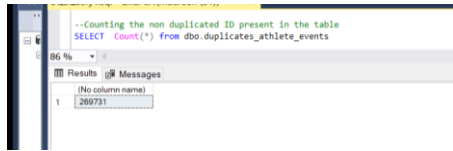
- a) Counting the number of rows in the dataset both athlete\_events and noc\_regions

Number of rows for the athlete events = 271116



- b) Removing the duplicate ID's by Creating a non duplicated table called duplicates\_athlete\_events

Number of rows retrieved for non duplicated row = 269731



c) in counting the non\_duplicate tables I discovered that the difference between the duplicate table and the non\_duplicate table is 1385, this means the athlete\_events has 1385 duplicated values

D) using a table to analyze the individual columns of the data

COLUMN ANALYZED	DATA TYPE	NOTES
ID	int	Duplicated ID = <b>271116</b> Non-duplicated ID = <b>269731</b> Total Number of Duplicate ID found = <b>1385</b>
NAME		
SEX	Int	Number of <b>females</b> that participated in the Olympics = <b>74378</b> Number of <b>Male</b> = <b>195353</b> Therefore, Male gets to participate in the Olympics than the female.
YEARS/AGE	int	Here athlete can participate in the Olympics more than once in different years E.g.: <b>ID 100046</b> at age 25 participated in the Olympics in year 1996 and at age 29 in year 2000  Calculating the birthdate of each athlete using the DATEADD function, I found out that there is no consistency in the birthdate of some athlete e.g.: Athlete <b>ID 100046</b> : Calculating the birthdate at age = 24 year = 2008, <b>birthdate = 1984</b> AND At age = 27, Year =2012

		<b>birthdate =1985.</b> This means that some of the athletes are not consistent with their age.
Height	int	
Weight	int	
Team	nvarchar	From the analysis here, an athlete can participate in different teams/ country during each Olympic year E.g., athlete <b>ID 122408</b> can be in <b>team Belarus</b> and <b>team Ukraine</b> at different Olympic year.
NOC	nvarchar	I figured out that same athlete ID can have different NOC i.e. same athlete can represent a 2 different country in different Olympics season, and this might be inconsistency of the data. E.g.: Example <b>ID = 87787</b> represented <b>Nigeria</b> for 2 Olympics season in the year = 1996 and 2000 but in the year 2004 and 2008 he represented <b>Portugal</b>
Sport	nvarchar	In this analysis same athlete ID can participate in different sport E.g.: <b>ID 1407</b> changed from <b>Rowing</b> to <b>Water Polo</b>
Games	nvarchar	This is a concatenation of <b>Year</b> and <b>Season</b>
Events		

**3). Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.**

The two-exploration data are:

#### **Athlete and Region**

It appears that 9653 Athlete from the USA region participated in the Olympics.

```

Select count (distinct ID) as num_athlete , region
into dbo.athlete_region
from dbo.BD_Events as E left join dbo.BD_Noc as N
on E.Noc = N.Noc
group by region
order by num_athlete desc

```

	num_athlete	region
1	9653	USA
2	7575	Germany
3	6281	UK
4	6170	France
5	5610	Russia
6	4935	Italy
7	4812	Canada
8	4067	Japan
9	3870	Australia
10	3787	Sweden
11	2985	China
12	2970	Poland
13	2939	Netherl...
14	2883	Switzer...
15	2782	Czech ...
16	2761	Hungary
17	2637	Spain
18	2393	South ...
19	2347	Finland
20	2337	Austria
21	2216	Norway
22	2078	Belgium

## Medals by Region

It appears the athlete ID 94406 has the highest number of medals won

```

Select
(E.ID),
F.Name,
N.Region,
Count(E.Medal) as Num_Medals
Into dbo.athlete_medals
from dbo.BD_Events as E left join dbo.BD_Noc as N
on E.Noc = N.Noc left join dbo.Fourth_athlete as F
on E.ID = F.ID
where Medal not in ('NA')
group by E.ID,F.Name,N.Region
order by Num_medals desc

```



Results				
	ID	Name	Region	Num_Medals
1	94406	Michael Fred Phelps, II	USA	28
2	67046	Larysa Semenivna Latynina (Diriy-)	Russia	18
3	4198	Nikolay Yefimovich Andrianov	Russia	15
4	11951	Ole Einar Bjørndalen	Norway	13
5	74420	Edoardo Mangiarotti	Italy	13
6	89187	Takashi Ono	Japan	13
7	109161	Borys Anfiyanovich Shakhlin	Russia	13
8	119922	Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	USA	12
9	121258	Dara Grace Torres (-Hoffman, -Minas)	USA	12
10	87390	Paavo Johannes Nurmi	Finland	12
11	85286	Aleksey Yuryevich Nemov	Russia	12
12	70965	Ryan Steven Lochte	USA	12
13	57998	Sawao Kato	Japan	12
14	23426	Natalie Anne Coughlin (-Hall)	USA	12
15	35550	Birgit Fischer-Schmidt	Germany	12
16	21402	Viktor Ivanovich Chukarin	Russia	11
17	18826	Vra slavsk (-Odloilov)	Czech Republic	11
18	11642	Matthew Nicholas "Matt" Biondi	USA	11

## Games with observations

The highest number of games was played in Sydney in games/ Olympics 2000 summer

```

Select E.Games, G.City,
count (E.ID) as Num_games_obs
into dbo.Num_games_obs
from dbo.BD_Events as E left join dbo.BD_Games_info as G
on E.ID = G.ID
group by E.Games, G.city
order by Num_games_obs desc

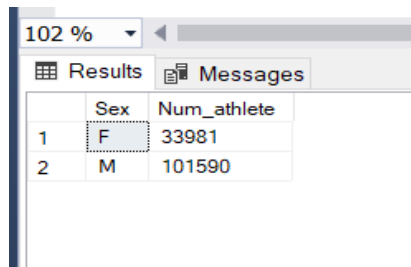
```

	Games	City	Num_games_obs
1	2000 Summer	Sydney	13821
2	1996 Summer	Atlanta	13780
3	2016 Summer	Rio de Janeiro	13688
4	2008 Summer	Beijing	13602
5	2004 Summer	Athina	13443
6	1992 Summer	Barcelona	12977
7	2012 Summer	London	12920
8	1988 Summer	Seoul	12037
9	1972 Summer	Munich	10304
10	1984 Summer	Los Angeles	9454
11	1976 Summer	Montreal	8641
12	1968 Summer	Mexico City	8588
13	1952 Summer	Helsinki	8270
14	1960 Summer	Roma	8119
15	1964 Summer	Tokyo	7702
16	1980 Summer	Moskva	7191
17	1936 Summer	Berlin	6506
18	1948 Summer	London	6408
19	1924 Summer	Paris	5249
20	1928 Summer	Amsterdam	4992
21	2014 Winter	Sochi	4891
22	2004 Summer	Beijing	4845
23	1956 Summer	Melbourne	4829
24	2000 Summer	Athina	4791
25	2008 Summer	Athina	4730

## Athlete by gender

Males are more present in the Olympics than female.

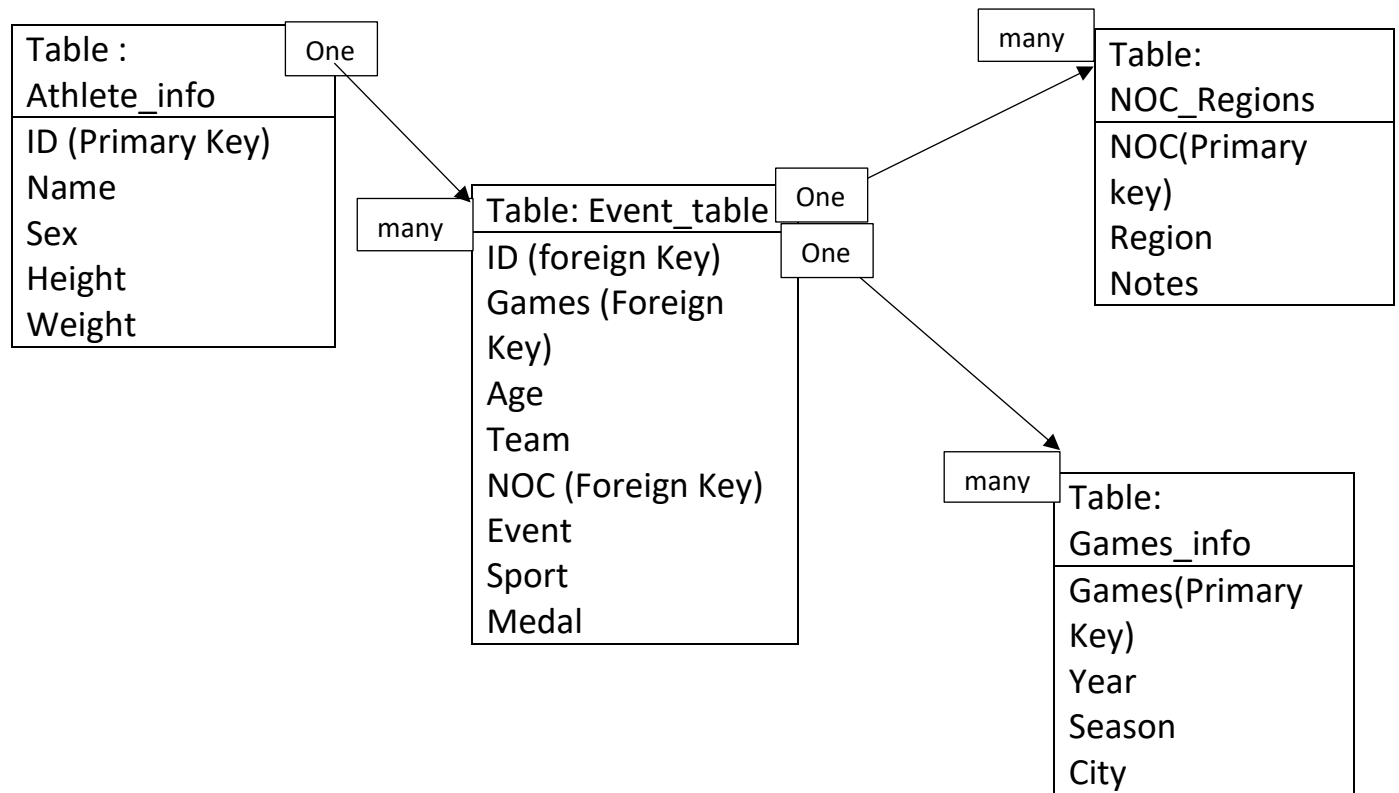
```
select Sex ,Count(ID) as Num_athlete  
into dbo.Gender  
from dbo.Fourth_athlete  
group by sex
```



The screenshot shows a SQL Server query results window. The window has a title bar with '102 %' and a scroll bar. Below the title bar are two tabs: 'Results' and 'Messages'. The 'Results' tab is active, displaying a table with two columns: 'Sex' and 'Num\_athlete'. The table contains two rows of data: one for females (F) with a count of 33981, and one for males (M) with a count of 101590.

	Sex	Num_athlete
1	F	33981
2	M	101590

4) Create an ERD or Proposed ERD to show the relationships of data you are exploring



## Step 2 Develop project proposal

My project goal is to learn more about the Olympoic data for the past 120 years and know the type of games that are played, the number of athlete that has participated , the number of Games played by each athlete , the number of medal won by each country or each athlete and the number of females and male who has participated in the olympics game for the past 120 years.

My findings might intrest some olympics agencies who which to know the number of medals an athlete has won , Governmnet, to know the number of Gold, Silver and Bronze won by its Country, the Journalist to know the report on the contry that has participated in the olympics and aslo the number of athlete that has participated in the olympics over the years.

## Questions

- 1) Which country has participated most in the olympics .
- 2) Which Gender has the heighest participation and from which country.
- 3) Which gender has the lowest male participant.
- 4) The number of age that is most represnted in the Oympics.
- 5) Which country has the won the heighest Gold for Gymnastics .

## Hypothesis

- 1) USA has participated in more olympics country than any other country
- 2) Male has participated more in the Olympics than the female .
- 3) South sudan is the least participated country in the olympics
- 4) Age 23 - 40 has been the most represnted age in the olympics.
- 5) Russia is better in gymanastic and has won more gold medals than any other country.

## Approach

In order to approve or disprove my hypothesis I will visualize my findings by creating bar chats with columns like sports, age and years. I will Explore the relationship between Medal regions and athlete to kow the number of country that has the highest number of medals and I will evaluate them by using a dashboard.