SQL FOR DATA SCIENCE CAPSTONE PROJECT

Assignment 2: Week 2/ Milestone 2: Descriptive stats

Name: Okonkwo Maureen

Client: Sports Stats

Data set: Olympics Dataset

Date: April 2022

Information from Milestone 1

COLUMN ANALYZED	DATA TYPE	NOTES
ID	int	Duplicated ID = **271116** Non-duplicated ID = **269731** Total Number of Duplicate ID found = **1385**
NAME		
SEX	Int	Number of **females** that participated in the Olympics = **74378** Number of **Male = 195353** 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.: **ID 100046** 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 **ID 100046:** Calculating the birthdate at age = 24 year = 2008, **birthdate = 1984** AND At age = 27, Year =2012 **birthdate =1985.** 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 **ID 122408** can be in **team Belarus** and **team Ukraine** 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 **ID = 87787** represented **Nigeria** for 2 Olympics season in the year = 1996 and 2000 but in the year 2004 and 2008 he represented **Portugal**
Sport	nvarchar	In this analysis same athlete ID can participate in different sport E.g.: **ID 1407** changed from **Rowing** to **Water Polo**
Games	nvarchar	This is a concatenation of **Year** and **Season**
Events		

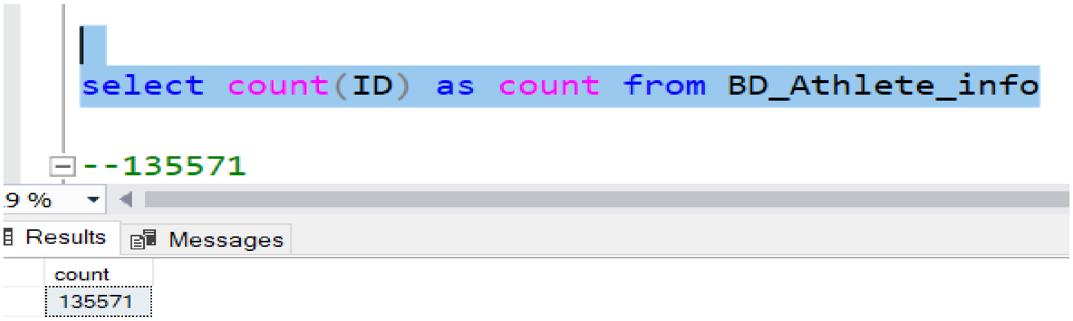
1) Summary of the different descriptive Statistics I used and which variables I used;

I looked at descriptive statistics of variables like Height, Weight, Sex, Age, medal, Events, sports and country.

In variables that has values like 'NA' I created a different table for 'NA' values and went ahead with the descriptive statistics of the various variables.

Column Name	Data Type	Row Name	Num of NA	Count	Max	Min	Avg
Height	nvarchar		33916	101655	226	127	176.32
		Female	-	30225	213	127	168.93
		Male	-	71430	226	127	179.44
Weight	nvarchar		34885	100686	99	100	76.466
		Female	-	29862	99	100	61.27
		Male	-	70824	99	100	76.46
Sex	nvachar		135571				
		Female	-	3381			
		Male	-	101590			

1) Using the **ATHLETE INFO TABLE** to extract our variables

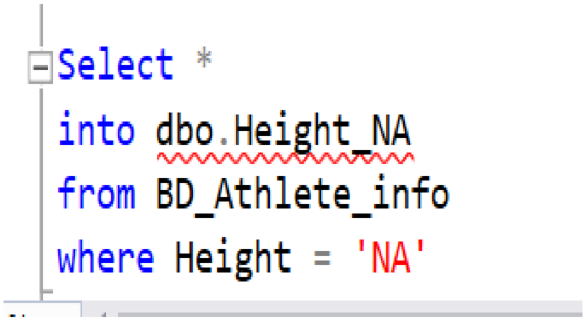


```sql
select count(ID) as count from BD_Athlete_info

--135571
```

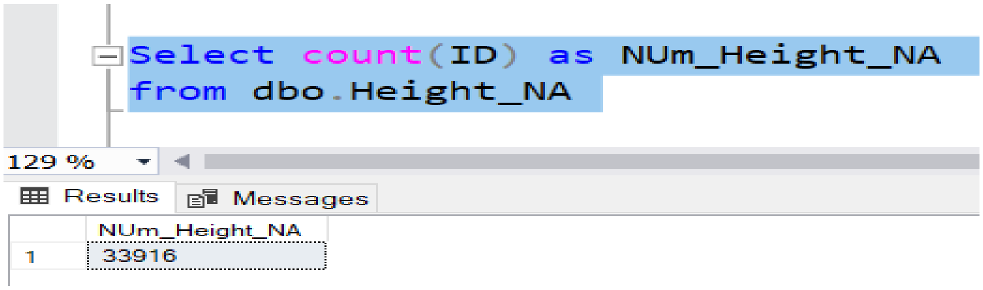
```
count
135571
```

2) Creating a table called **Height_NA** to computes all the height with value NA



```sql
Select *
into dbo.Height_NA
from BD_Athlete_info
where Height = 'NA'
```

3) Counting the number of NA values in the height variable.

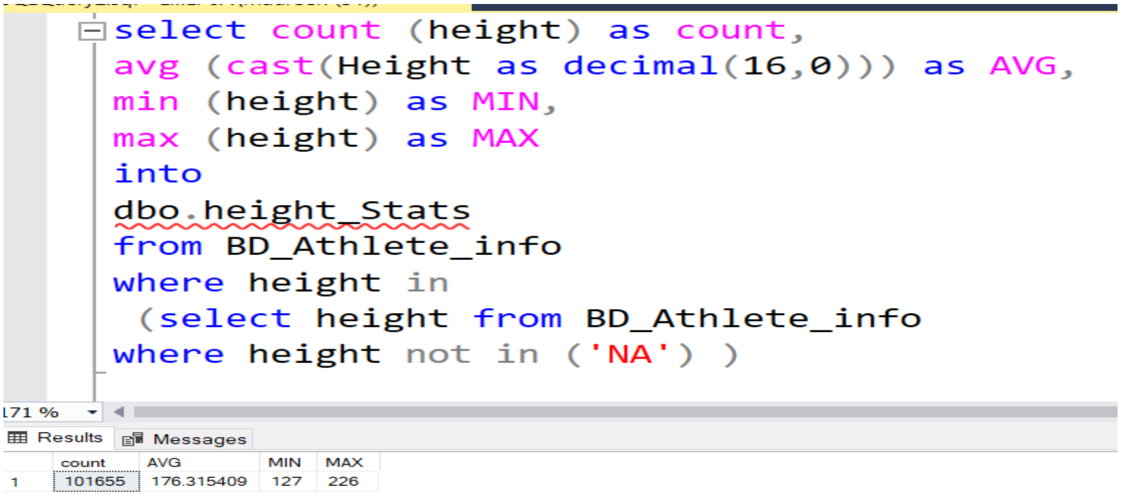


```sql
Select count(ID) as NUm_Height_NA
from dbo.Height_NA
```

```
   NUm_Height_NA
1  33916
```

4) Descriptive Statistics without NA

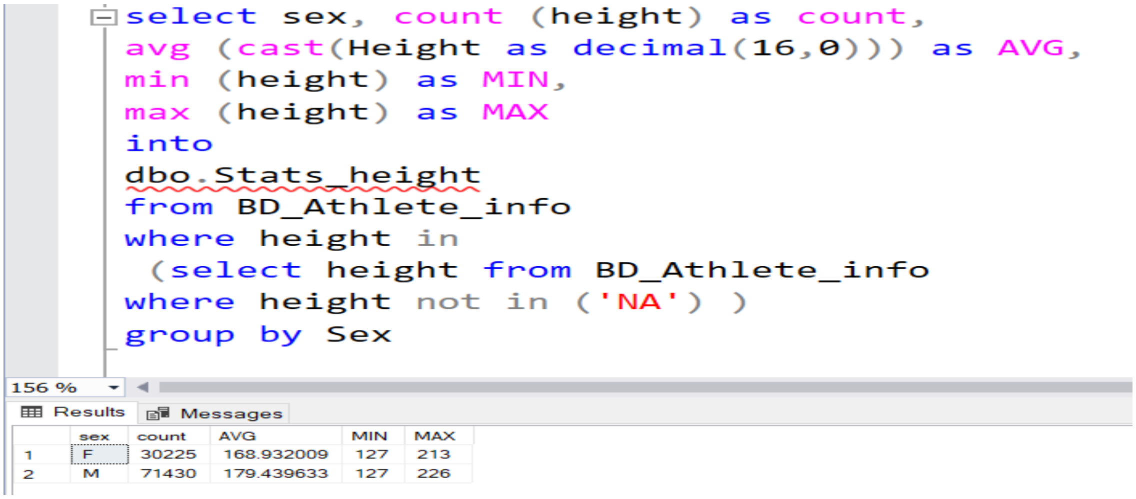
Variable Height:



```
select count (height) as count,
avg (cast(Height as decimal(16,0))) as AVG,
min (height) as MIN,
max (height) as MAX
into
dbo.height_Stats
from BD_Athlete_info
where height in
 (select height from BD_Athlete_info
where height not in ('NA') )
```

```
   count    AVG         MIN   MAX
1  101655   176.315409  127   226
```

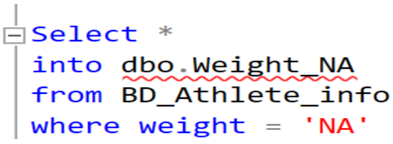
5) Height by Sex:



```
select sex, count (height) as count,
avg (cast(Height as decimal(16,0))) as AVG,
min (height) as MIN,
max (height) as MAX
into
dbo.Stats_height
from BD_Athlete_info
where height in
 (select height from BD_Athlete_info
where height not in ('NA') )
group by Sex
```

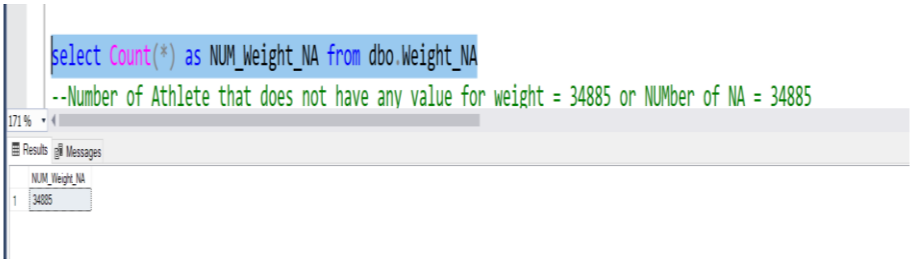
```
   sex  count  AVG         MIN  MAX
1  F    30225  168.932009  127  213
2  M    71430  179.439633  127  226
```

6) Creating a table called **Weight_NA** to computes all the height with value NA



```
Select *
into dbo.Weight_NA
from BD_Athlete_info
where weight = 'NA'
```

8) Counting the number of NA values in the Weight variable

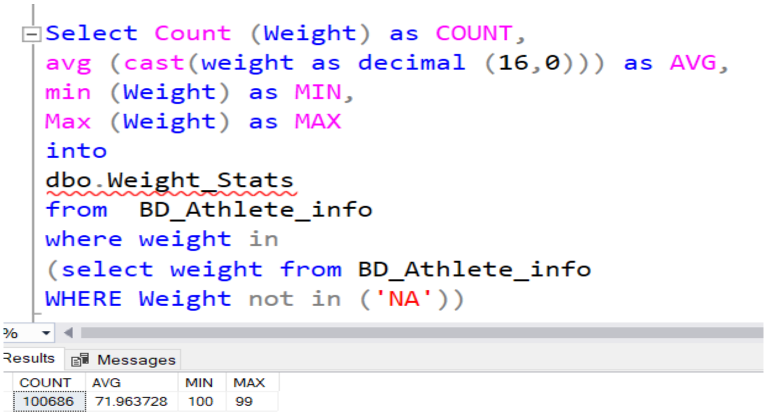


```
select Count(*) as NUM_Weight_NA from dbo.Weight_NA
--Number of Athlete that does not have any value for weight = 34885 or NUMber of NA = 34885
```

```
171 %
Results  Messages
   NUM_Weight_NA
1  34885
```

9) Descriptive Statistics without the NA for

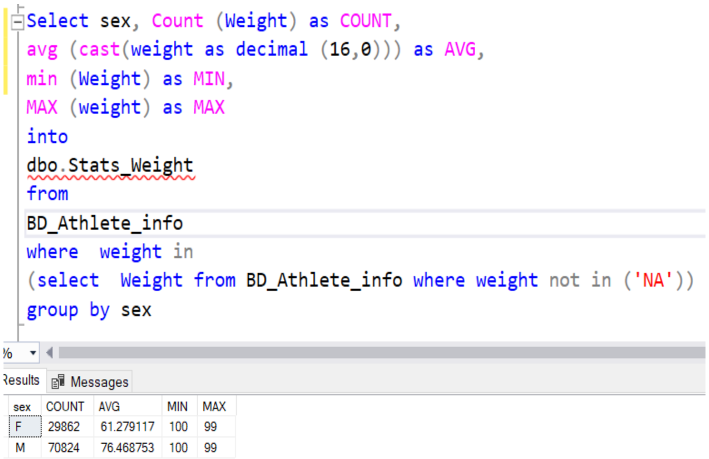
Variable Weight:



```
Select Count (Weight) as COUNT,
avg (cast(weight as decimal (16,0))) as AVG,
min (Weight) as MIN,
Max (Weight) as MAX
into
dbo.Weight_Stats
from  BD_Athlete_info
where weight in
(select weight from BD_Athlete_info
WHERE Weight not in ('NA'))
```

```
Results  Messages
COUNT    AVG        MIN   MAX
100686   71.963728  100   99
```

Weight by Sex:



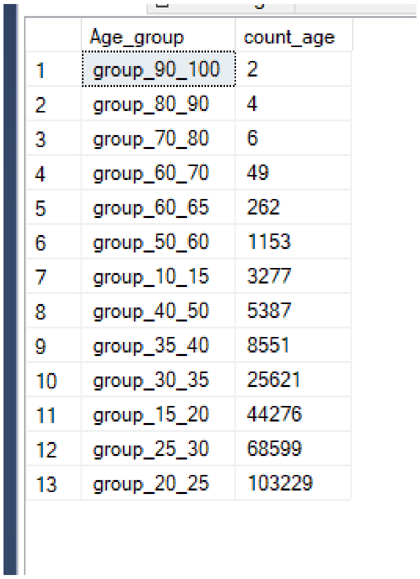
```
Select sex, Count (Weight) as COUNT,
avg (cast(weight as decimal (16,0))) as AVG,
min (Weight) as MIN,
MAX (weight) as MAX
into
dbo.Stats_Weight
from
BD_Athlete_info
where  weight in
(select  Weight from BD_Athlete_info where weight not in ('NA'))
group by sex
```

```
Results  Messages
sex  COUNT  AVG        MIN  MAX
F    29862  61.279117  100  99
M    70824  76.468753  100  99
```

Table: Event_Table

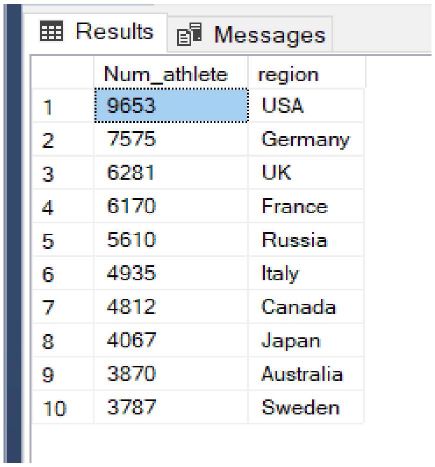
Column Name	Data Type	Number of activities	Number of NA	Count	Avg	Min	Max
Games	nvarchar	Total of 51 games played in the Olympics					
Age	nvarchar		9315	260416	25.454	10	97
Team							
NOC							
Event		765 events					
Sport		66 Sports					
Medals		NA = 229959 Gold = 13369 Silver = 13108 Bronze = 13295					

Age by groups



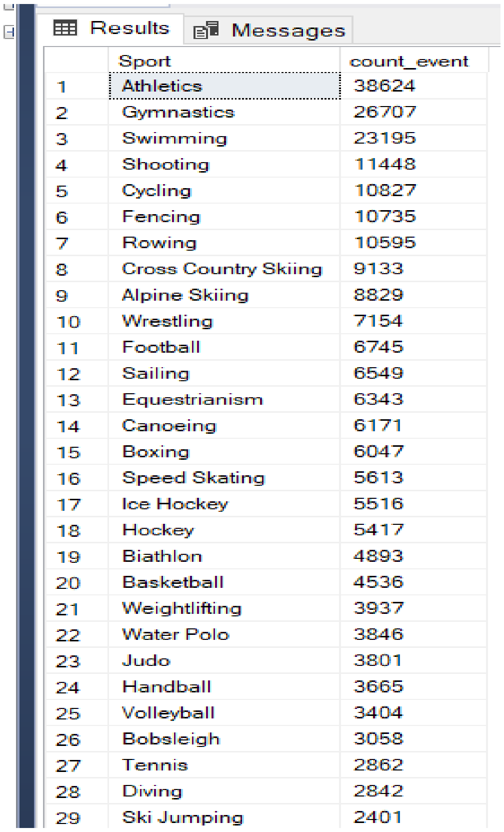
	Age_group	count_age
1	group_90_100	2
2	group_80_90	4
3	group_70_80	6
4	group_60_70	49
5	group_60_65	262
6	group_50_60	1153
7	group_10_15	3277
8	group_40_50	5387
9	group_35_40	8551
10	group_30_35	25621
11	group_15_20	44276
12	group_25_30	68599
13	group_20_25	103229

Atheletes by NOC Region (Country) - Retrieving the top 10



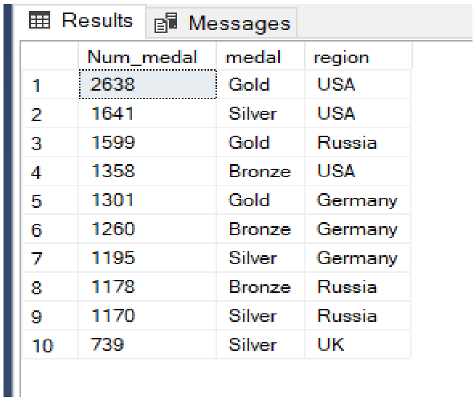
	Num_athlete	region
1	9653	USA
2	7575	Germany
3	6281	UK
4	6170	France
5	5610	Russia
6	4935	Italy
7	4812	Canada
8	4067	Japan
9	3870	Australia
10	3787	Sweden

Sport by events



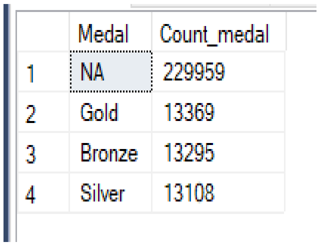
	Sport	count_event
1	Athletics	38624
2	Gymnastics	26707
3	Swimming	23195
4	Shooting	11448
5	Cycling	10827
6	Fencing	10735
7	Rowing	10595
8	Cross Country Skiing	9133
9	Alpine Skiing	8829
10	Wrestling	7154
11	Football	6745
12	Sailing	6549
13	Equestrianism	6343
14	Canoeing	6171
15	Boxing	6047
16	Speed Skating	5613
17	Ice Hockey	5516
18	Hockey	5417
19	Biathlon	4893
20	Basketball	4536
21	Weightlifting	3937
22	Water Polo	3846
23	Judo	3801
24	Handball	3665
25	Volleyball	3404
26	Bobsleigh	3058
27	Tennis	2862
28	Diving	2842
29	Ski Jumping	2401

Medals by NOC_region (Country)



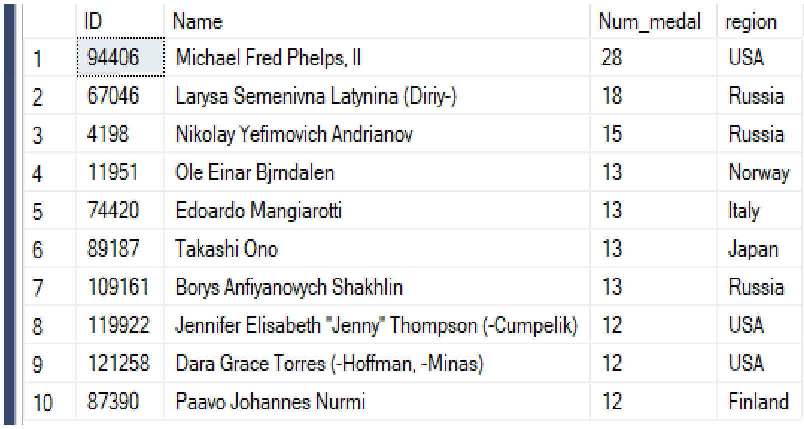
	Num_medal	medal	region
1	2638	Gold	USA
2	1641	Silver	USA
3	1599	Gold	Russia
4	1358	Bronze	USA
5	1301	Gold	Germany
6	1260	Bronze	Germany
7	1195	Silver	Germany
8	1178	Bronze	Russia
9	1170	Silver	Russia
10	739	Silver	UK

Total Number of medals won in the Olympics



	Medal	Count_medal
1	NA	229959
2	Gold	13369
3	Bronze	13295
4	Silver	13108

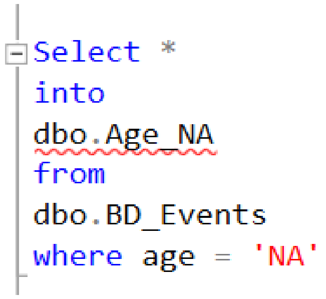
Top 10 athlete with the highest gold medal



	ID	Name	Num_medal	region
1	94406	Michael Fred Phelps, II	28	USA
2	67046	Larysa Semenivna Latynina (Diriy-)	18	Russia
3	4198	Nikolay Yefimovich Andrianov	15	Russia
4	11951	Ole Einar Bjrndalen	13	Norway
5	74420	Edoardo Mangiarotti	13	Italy
6	89187	Takashi Ono	13	Japan
7	109161	Borys Anfiyanovych Shakhlin	13	Russia
8	119922	Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	12	USA
9	121258	Dara Grace Torres (-Hoffman, -Minas)	12	USA
10	87390	Paavo Johannes Nurmi	12	Finland

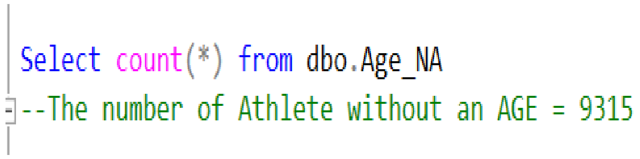
/Queries/

Variable: Age

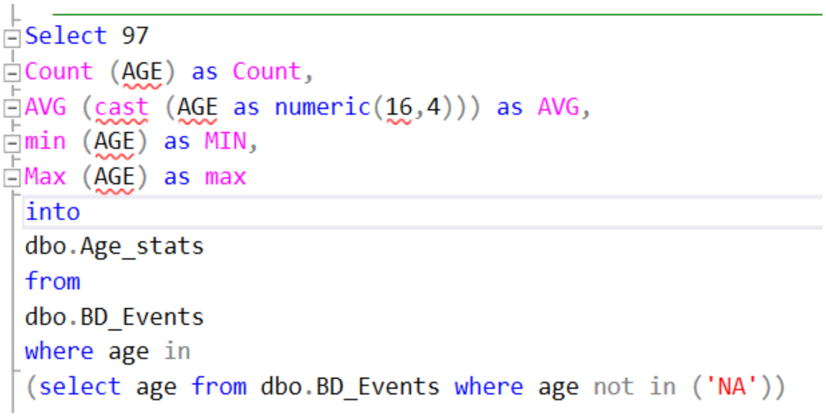


```
Select *
into
dbo.Age_NA
from
dbo.BD_Events
where age = 'NA'
```

Count Age

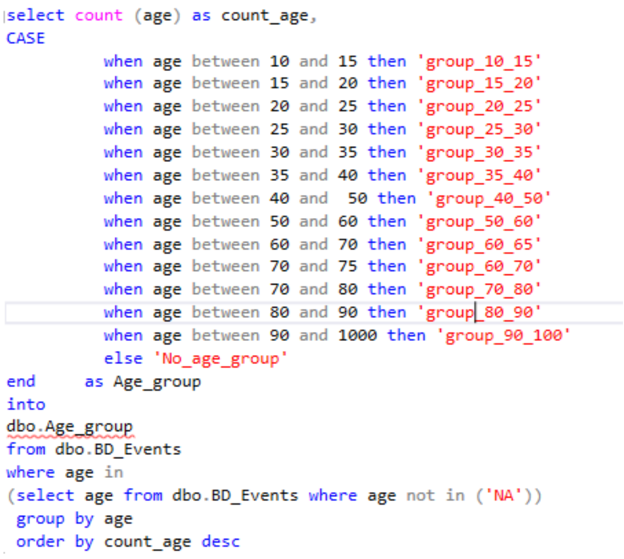


```
Select count(*) from dbo.Age_NA
--The number of Athlete without an AGE = 9315
```



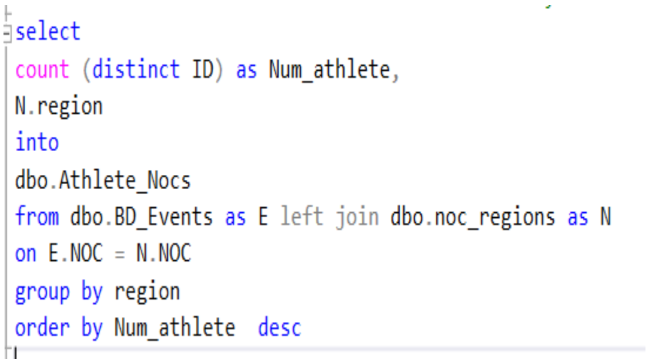
```
Select 97
Count (AGE) as Count,
AVG (cast (AGE as numeric(16,4))) as AVG,
min (AGE) as MIN,
Max (AGE) as max
into
dbo.Age_stats
from
dbo.BD_Events
where age in
(select age from dbo.BD_Events where age not in ('NA'))
```

Grouping Age by their age group



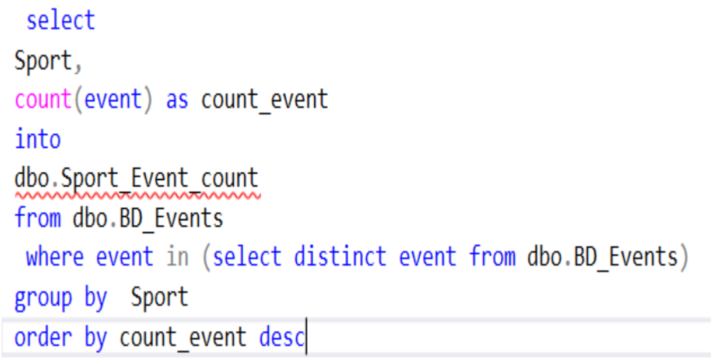
```
select count (age) as count_age,
CASE
          when age between 10 and 15 then 'group_10_15'
          when age between 15 and 20 then 'group_15_20'
          when age between 20 and 25 then 'group_20_25'
          when age between 25 and 30 then 'group_25_30'
          when age between 30 and 35 then 'group_30_35'
          when age between 35 and 40 then 'group_35_40'
          when age between 40 and  50 then 'group_40_50'
          when age between 50 and 60 then 'group_50_60'
          when age between 60 and 70 then 'group_60_65'
          when age between 70 and 75 then 'group_60_70'
          when age between 70 and 80 then 'group_70_80'
          when age between 80 and 90 then 'group_80_90'
          when age between 90 and 1000 then 'group_90_100'
          else 'No_age_group'
end     as Age_group
into
dbo.Age_group
from dbo.BD_Events
where age in
(select age from dbo.BD_Events where age not in ('NA'))
 group by age
 order by count_age desc
```

Variable: Athlete and NOC



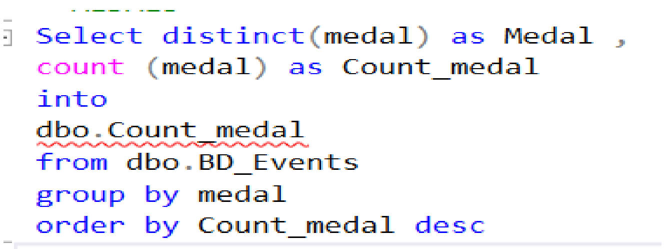
```
select
count (distinct ID) as Num_athlete,
N.region
into
dbo.Athlete_Nocs
from dbo.BD_Events as E left join dbo.noc_regions as N
on E.NOC = N.NOC
group by region
order by Num_athlete  desc
```

Variable: Sport by the number Events



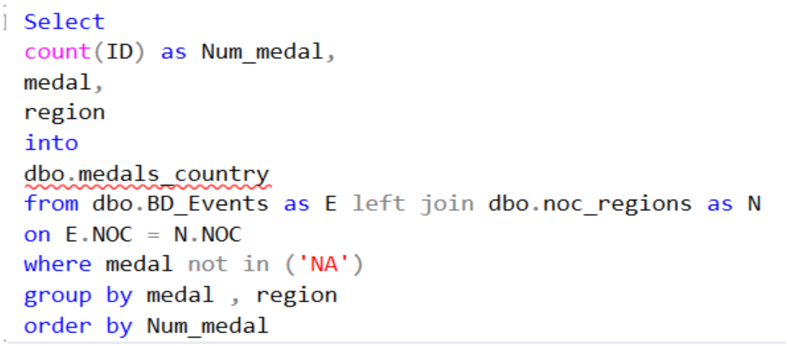
```
 select
Sport,
count(event) as count_event
into
dbo.Sport_Event_count
from dbo.BD_Events
 where event in (select distinct event from dbo.BD_Events)
group by  Sport
order by count_event desc
```

Variable: Medals



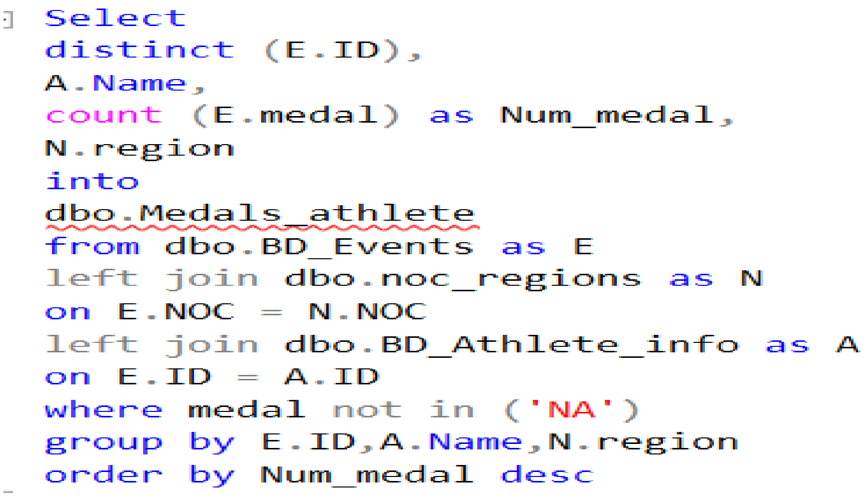
```
Select distinct(medal) as Medal ,
count (medal) as Count_medal
into
dbo.Count_medal
from dbo.BD_Events
group by medal
order by Count_medal desc
```

Variable: Medal by NOC (Region or Country)



```
Select
count(ID) as Num_medal,
medal,
region
into
dbo.medals_country
from dbo.BD_Events as E left join dbo.noc_regions as N
on E.NOC = N.NOC
where medal not in ('NA')
group by medal , region
order by Num_medal
```

Variable: Athlete with the highest number of medals in the Olympic history;



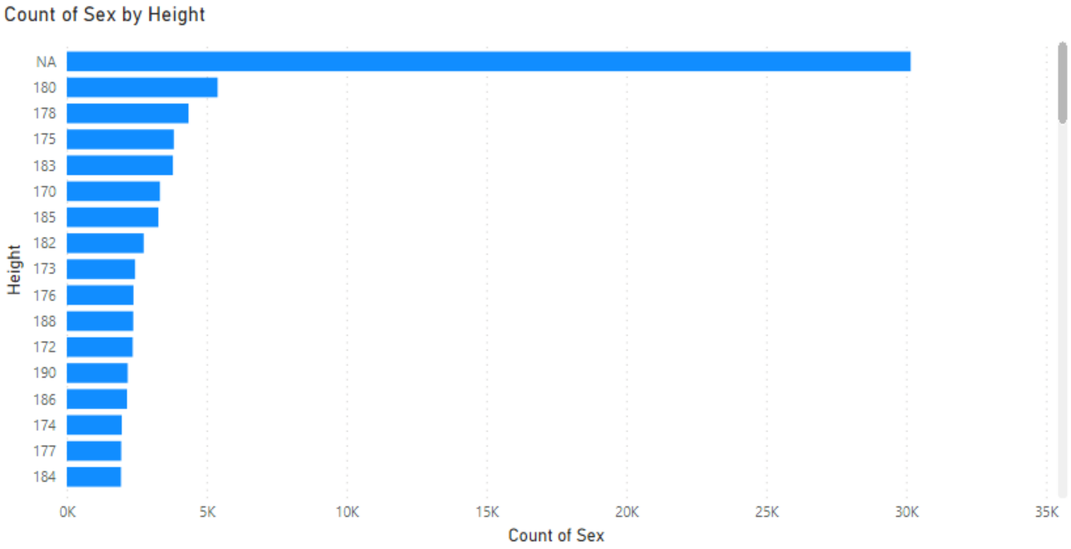
```
Select
distinct (E.ID),
A.Name,
count (E.medal) as Num_medal,
N.region
into
dbo.Medals_athlete
from dbo.BD_Events as E
left join dbo.noc_regions as N
on E.NOC = N.NOC
left join dbo.BD_Athlete_info as A
on E.ID = A.ID
where medal not in ('NA')
group by E.ID,A.Name,N.region
order by Num_medal desc
```

2) Submit 2-3 key points you may have discovered about the data e.g., new relationships,

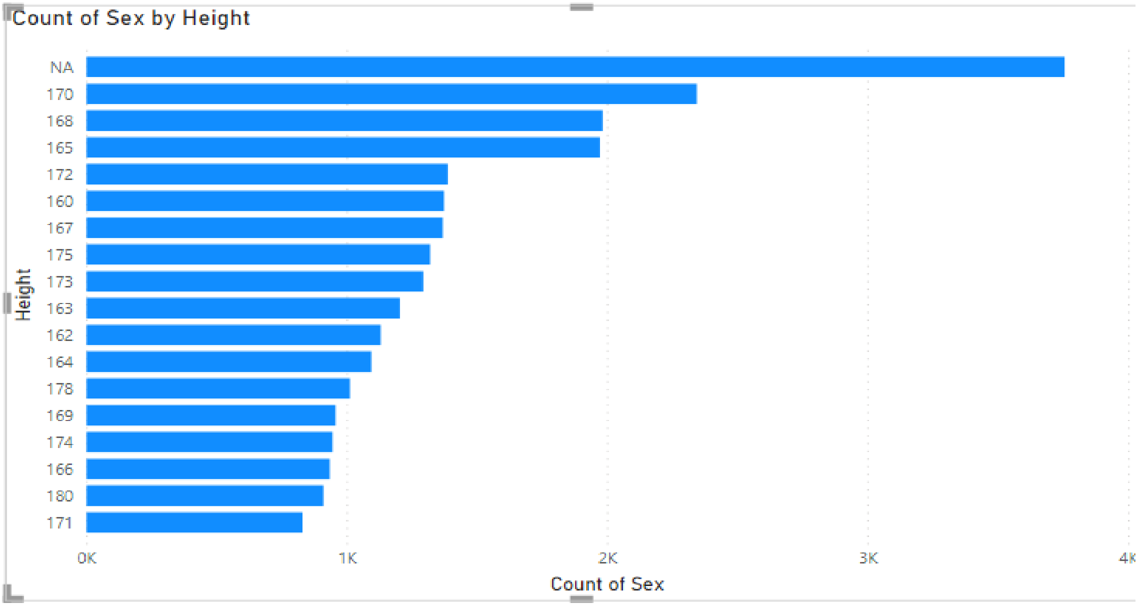
Did you come up with additional ideas for other things to review.

A) The variable Sex: Height: There are more than twice the males than the females, and as expected the average height of the females is lower than the average height of the males. But the surprising thing is both the male and the female minimum and maximum height are almost the same.

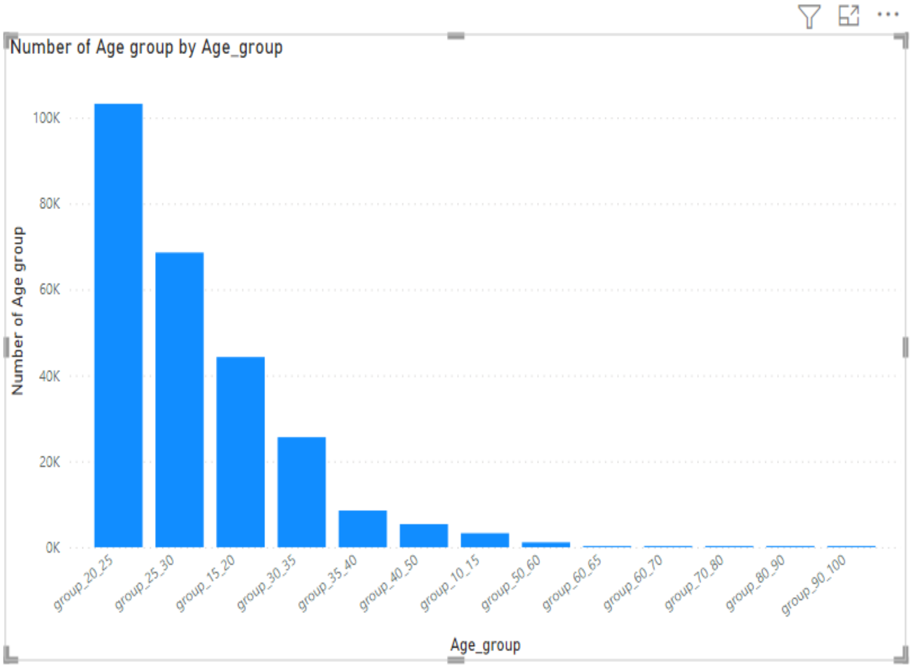
Male Height Stats



Female Height Stats

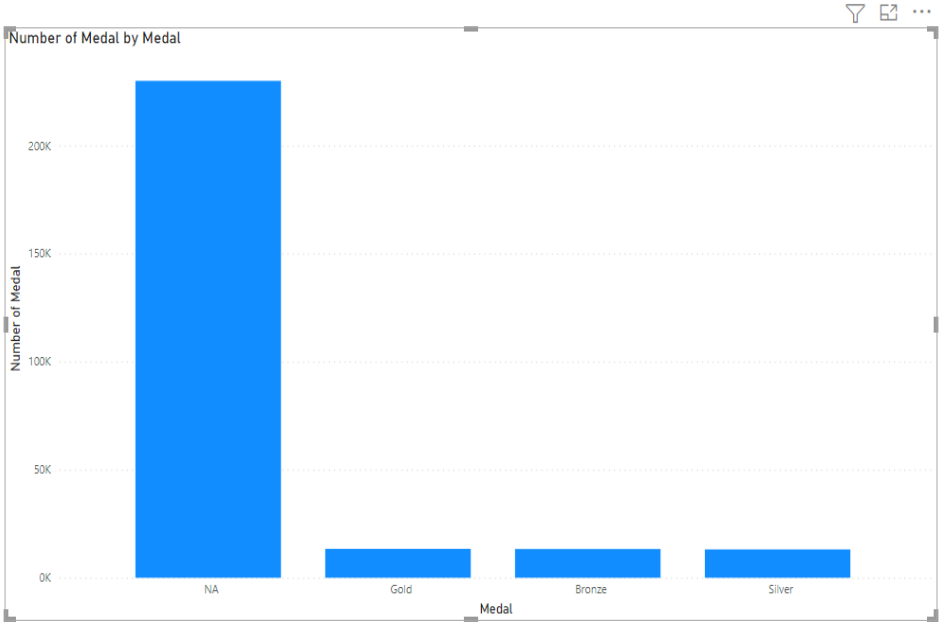


B) Variable : Age : Ages Between 15- 50 are well represned but the most predominated are age between group 20 – 25 years old



C) Variable Country (NOC): The USA is the country with the most participation in the Olympics

D) Variable Medal: The number of participants that have not won a medal is quite high about 15% participant won either Gold, Silver or Bronze and the top 4 countries that has won the highest number medals are USA, Russia, Germany and UK .



3) Did you prove or disprove any of your initial hypothesis? If so, which one and what do you plan to do next

Hypothesis

1) The age group 20-25 is the most participated age in the Olympics (proved)
2) Athletics has more participants than any other sports (Proved)
3) The athlete Michael Fred Phelps, II from USA is the athlete with the highest number of gold medal which is 23 gold medal and also the athlete with the highest number of medals which is 28 medals in total.
4) Russia has won the highest gold medal in Gymnastic competition, which makes them best in gymnastics competition while the USA is best in swimming as they have won the highest gold medal in swimming competition

4) What additional questions are you seeking?

1) Does height have any effect to the performance of some athlete and is height really an advantage in some sports.
2) Does gender (Sex) have anything to do with the number of medals to won.
3) Does choice of sport have anything to do with the number of medals to won.