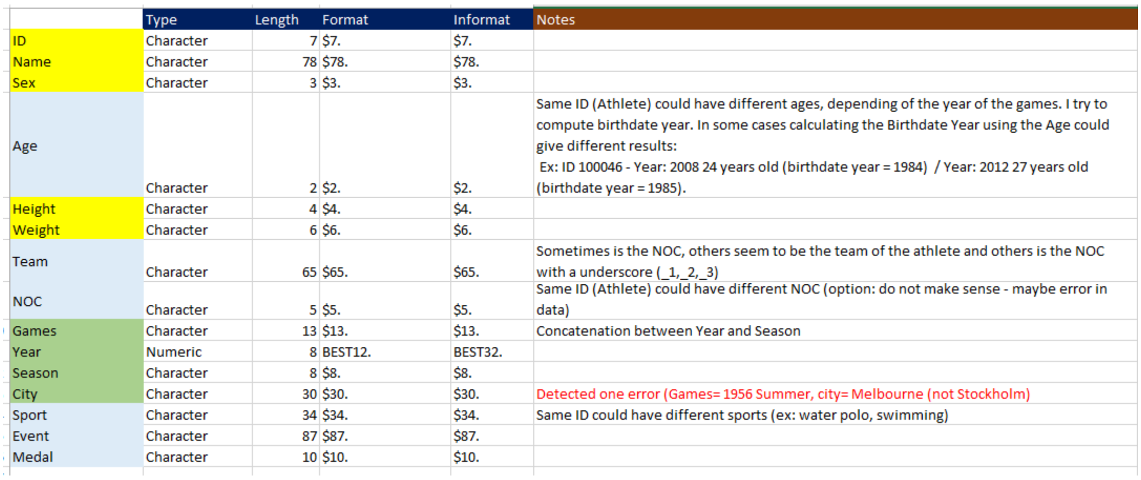
SQL for Data Science Capstone Project

Assignment 2: WEEK 2 / MILSTONE 2: Descriptive Stats

Name: Daniel Siva

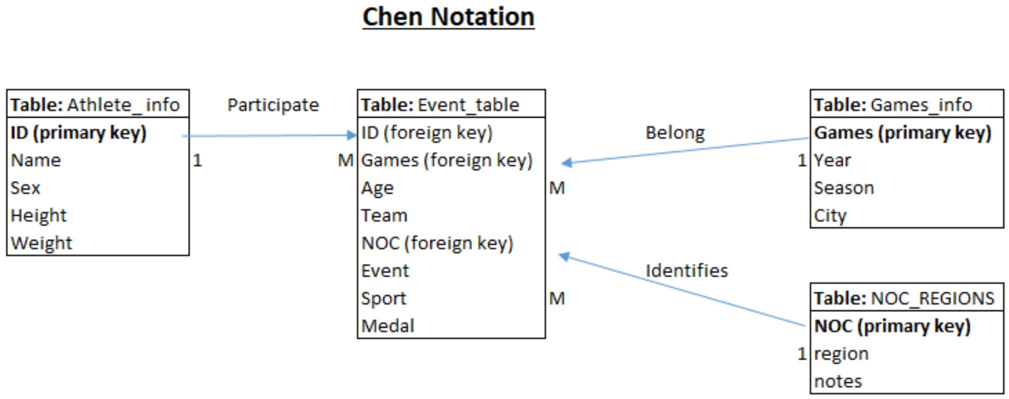
0) Information of week1:

After analyze the dataset I decided to separate the main information in 3 tables: ‘Athlete_info’, ‘Event_Table’, ‘Games_info’ and joined the table ‘NOC_REGIONS’ (see the EDR for more information).



	Type	Length	Format	Informat	Notes
ID	Character	7	$7.	$7.	
Name	Character	78	$78.	$78.	
Sex	Character	3	$3.	$3.	
Age	Character	2	$2.	$2.	Same ID (Athlete) could have different ages, depending of the year of the games. I try to compute birthdate year. In some cases calculating the Birthdate Year using the Age could give different results: Ex: ID 100046 - Year: 2008 24 years old (birthdate year = 1984) / Year: 2012 27 years old (birthdate year = 1985).
Height	Character	4	$4.	$4.	
Weight	Character	6	$6.	$6.	
Team	Character	65	$65.	$65.	Sometimes is the NOC, others seem to be the team of the athlete and others is the NOC with a underscore (_1,_2,_3)
NOC	Character	5	$5.	$5.	Same ID (Athlete) could have different NOC (option: do not make sense - maybe error in data)
Games	Character	13	$13.	$13.	Concatenation between Year and Season
Year	Numeric	8	BEST12.	BEST32.	
Season	Character	8	$8.	$8.	
City	Character	30	$30.	$30.	Detected one error (Games= 1956 Summer, city= Melbourne (not Stockholm)
Sport	Character	34	$34.	$34.	Same ID could have different sports (ex: water polo, swimming)
Event	Character	87	$87.	$87.	
Medal	Character	10	$10.	$10.	

Entity Relationship Diagram

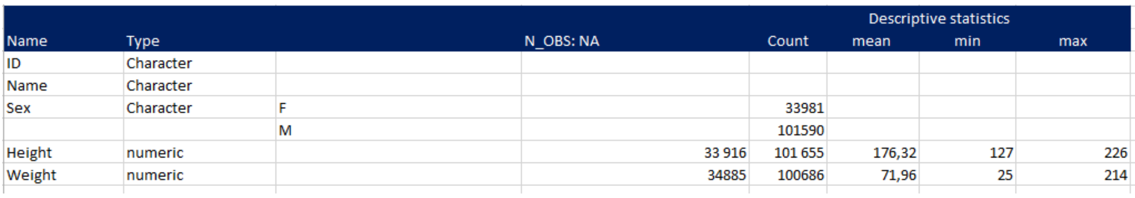


1) Provide a summary of the different descriptive statistics you looked at and why.

I looked and compute some statistics for each table. One of my concerns was about the 'NA' value. If the data about the variable is not available I should not use the observation to make assumptions. If the number of 'NA' is high maybe I need to change my analysis.

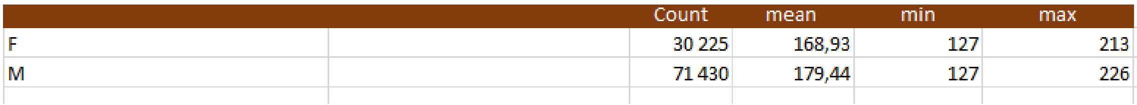
I focused my analysis in country, gender (sex), age, sports and medals to understand better this variables that are important to prove my hypothesis.

Table: Athlete_info



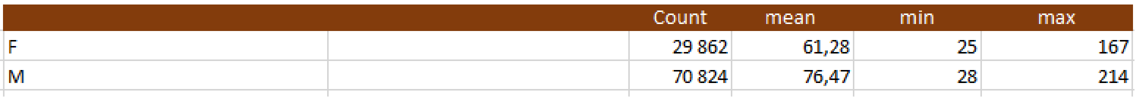
					Descriptive statistics		
Name	Type		N_OBS: NA	Count	mean	min	max
ID	Character						
Name	Character						
Sex	Character	F		33981			
		M		101590			
Height	numeric		33 916	101 655	176,32	127	226
Weight	numeric		34885	100686	71,96	25	214

Height by sex



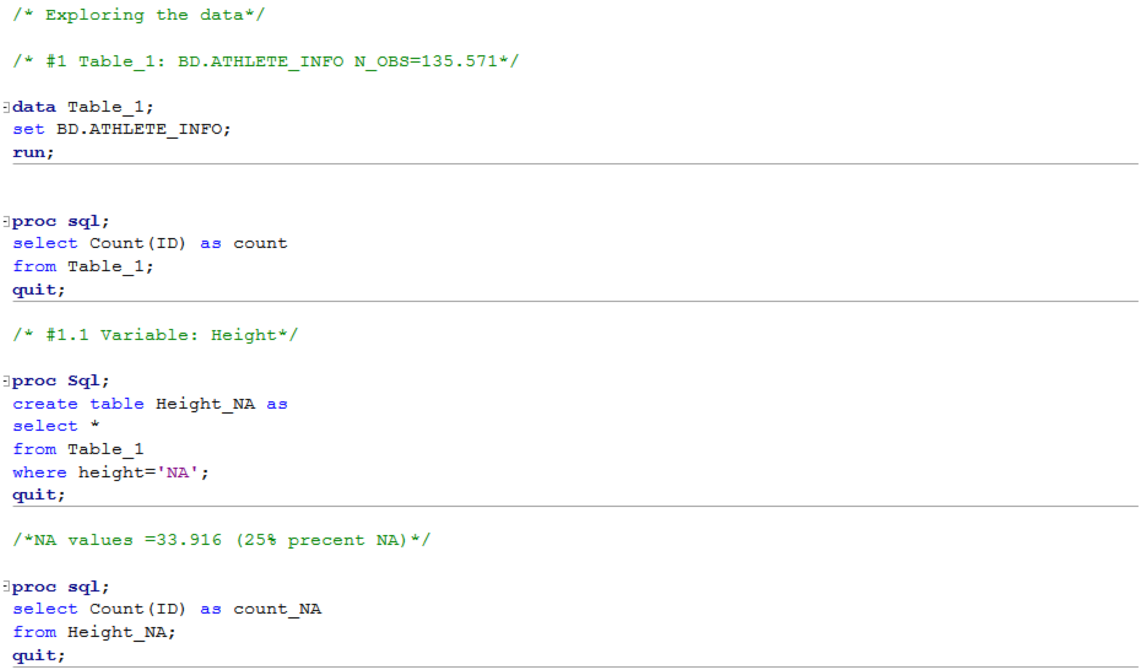
		Count	mean	min	max
F		30 225	168,93	127	213
M		71 430	179,44	127	226

Weight by sex

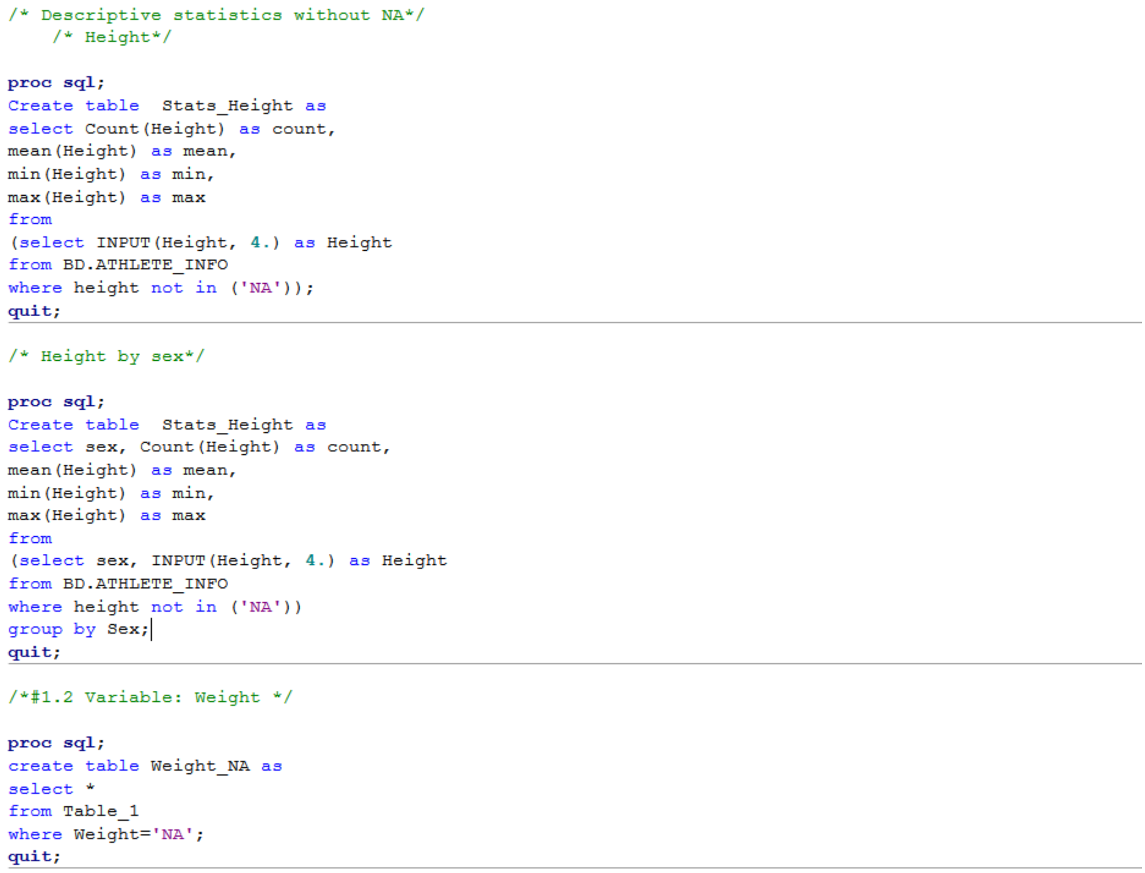


		Count	mean	min	max
F		29 862	61,28	25	167
M		70 824	76,47	28	214

Code (SAS using proc sql) – some examples:



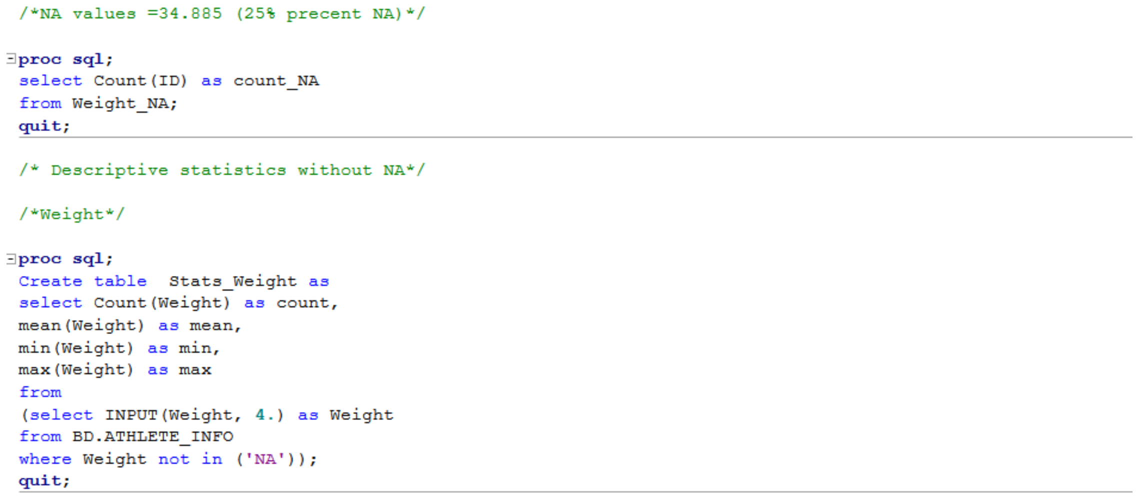
```
/* Exploring the data*/

/* #1 Table_1: BD.ATHLETE_INFO N_OBS=135.571*/

data Table_1;
set BD.ATHLETE_INFO;
run;


proc sql;
select Count(ID) as count
from Table_1;
quit;

/* #1.1 Variable: Height*/

proc Sql;
create table Height_NA as
select *
from Table_1
where height='NA';
quit;

/*NA values =33.916 (25% precent NA)*/

proc sql;
select Count(ID) as count_NA
from Height_NA;
quit;
```



```
/* Descriptive statistics without NA*/
    /* Height*/

proc sql;
Create table  Stats_Height as
select Count(Height) as count,
mean(Height) as mean,
min(Height) as min,
max(Height) as max
from
(select INPUT(Height, 4.) as Height
from BD.ATHLETE_INFO
where height not in ('NA'));
quit;
```

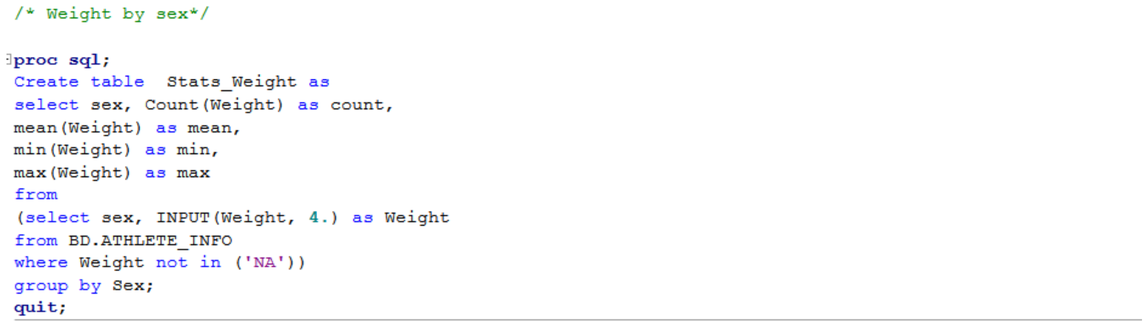
```
/* Height by sex*/

proc sql;
Create table  Stats_Height as
select sex, Count(Height) as count,
mean(Height) as mean,
min(Height) as min,
max(Height) as max
from
(select sex, INPUT(Height, 4.) as Height
from BD.ATHLETE_INFO
where height not in ('NA'))
group by Sex;
quit;
```

```
/*#1.2 Variable: Weight */

proc sql;
create table Weight_NA as
select *
from Table_1
where Weight='NA';
quit;
```



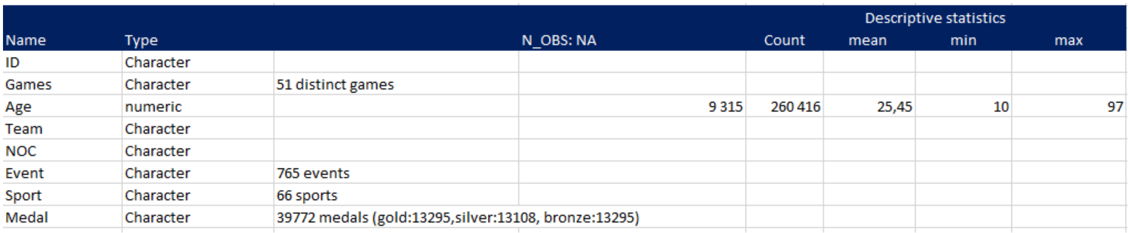
```
/*NA values =34.885 (25% precent NA)*/

proc sql;
select Count(ID) as count_NA
from Weight_NA;
quit;
```

```
/* Descriptive statistics without NA*/

/*Weight*/

proc sql;
Create table  Stats_Weight as
select Count(Weight) as count,
mean(Weight) as mean,
min(Weight) as min,
max(Weight) as max
from
(select INPUT(Weight, 4.) as Weight
from BD.ATHLETE_INFO
where Weight not in ('NA'));
quit;
```



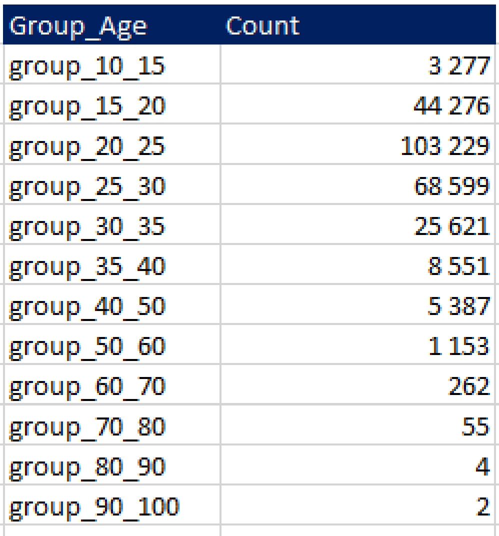
```
/* Weight by sex*/

proc sql;
Create table  Stats_Weight as
select sex, Count(Weight) as count,
mean(Weight) as mean,
min(Weight) as min,
max(Weight) as max
from
(select sex, INPUT(Weight, 4.) as Weight
from BD.ATHLETE_INFO
where Weight not in ('NA'))
group by Sex;
quit;
```

Table: Event_table



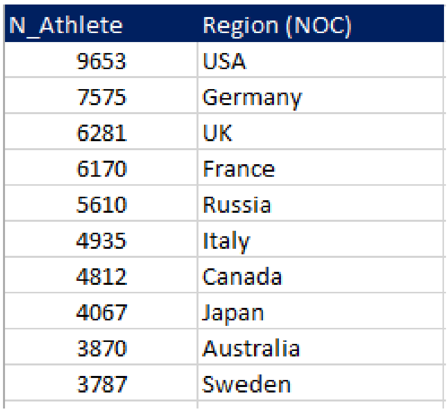
					Descriptive statistics		
Name	Type		N_OBS: NA	Count	mean	min	max
ID	Character						
Games	Character	51 distinct games					
Age	numeric		9 315	260 416	25,45	10	97
Team	Character						
NOC	Character						
Event	Character	765 events					
Sport	Character	66 sports					
Medal	Character	39772 medals (gold:13295,silver:13108, bronze:13295)					

Age by groups



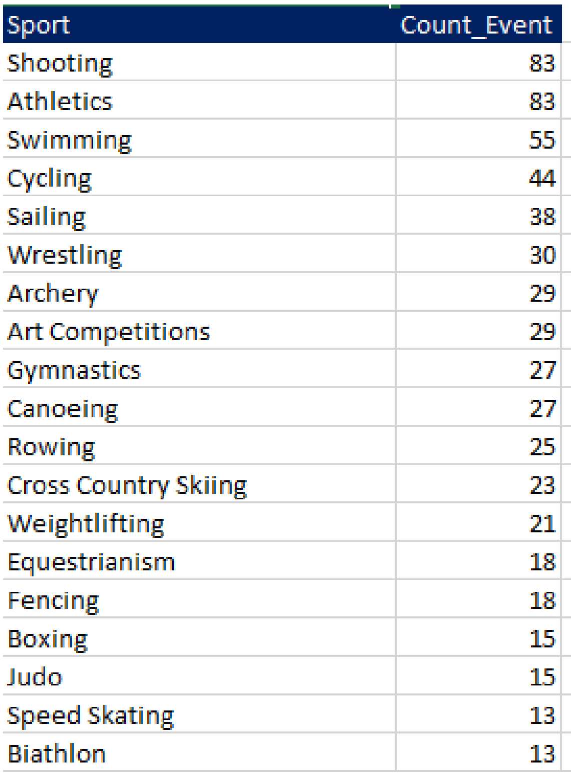
Group_Age	Count
group_10_15	3 277
group_15_20	44 276
group_20_25	103 229
group_25_30	68 599
group_30_35	25 621
group_35_40	8 551
group_40_50	5 387
group_50_60	1 153
group_60_70	262
group_70_80	55
group_80_90	4
group_90_100	2

Athletes by NOC (country) – top 10



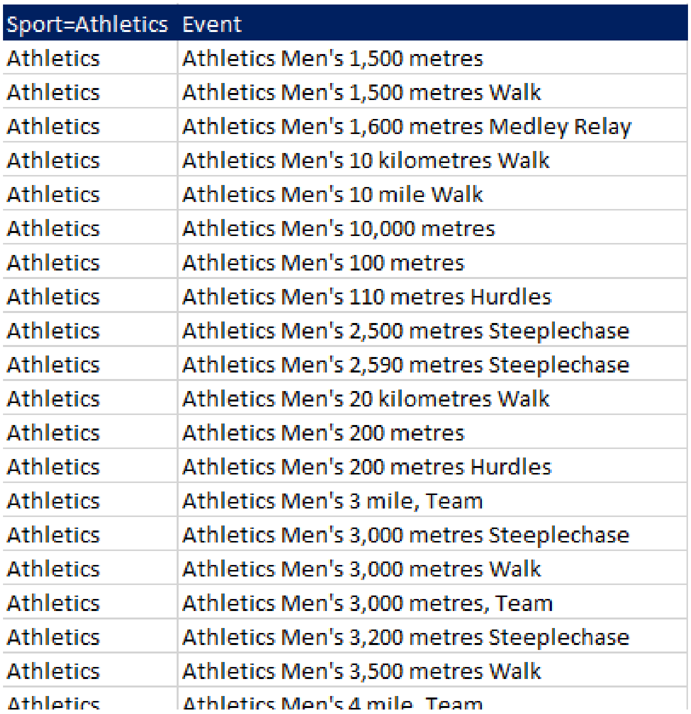
N_Athlete	Region (NOC)
9653	USA
7575	Germany
6281	UK
6170	France
5610	Russia
4935	Italy
4812	Canada
4067	Japan
3870	Australia
3787	Sweden

Sport (total 66) by event (example)



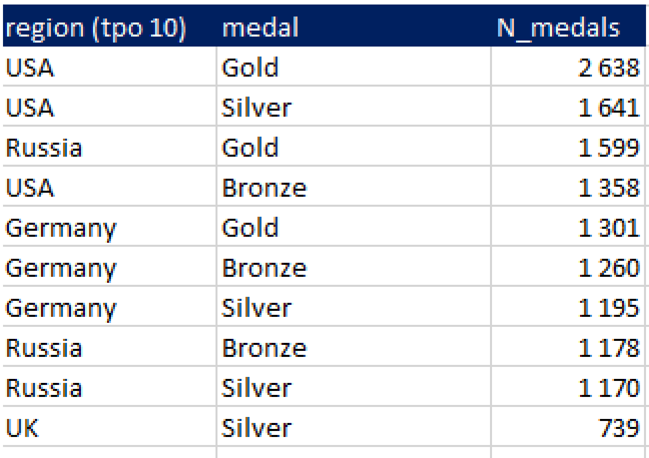
Sport	Count_Event
Shooting	83
Athletics	83
Swimming	55
Cycling	44
Sailing	38
Wrestling	30
Archery	29
Art Competitions	29
Gymnastics	27
Canoeing	27
Rowing	25
Cross Country Skiing	23
Weightlifting	21
Equestrianism	18
Fencing	18
Boxing	15
Judo	15
Speed Skating	13
Biathlon	13

Sport=Athletics events (83 example)



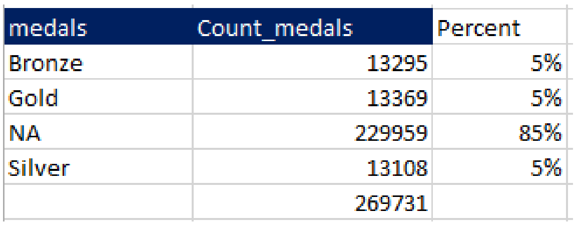
Sport=Athletics	Event
Athletics	Athletics Men's 1,500 metres
Athletics	Athletics Men's 1,500 metres Walk
Athletics	Athletics Men's 1,600 metres Medley Relay
Athletics	Athletics Men's 10 kilometres Walk
Athletics	Athletics Men's 10 mile Walk
Athletics	Athletics Men's 10,000 metres
Athletics	Athletics Men's 100 metres
Athletics	Athletics Men's 110 metres Hurdles
Athletics	Athletics Men's 2,500 metres Steeplechase
Athletics	Athletics Men's 2,590 metres Steeplechase
Athletics	Athletics Men's 20 kilometres Walk
Athletics	Athletics Men's 200 metres
Athletics	Athletics Men's 200 metres Hurdles
Athletics	Athletics Men's 3 mile, Team
Athletics	Athletics Men's 3,000 metres Steeplechase
Athletics	Athletics Men's 3,000 metres Walk
Athletics	Athletics Men's 3,000 metres, Team
Athletics	Athletics Men's 3,200 metres Steeplechase
Athletics	Athletics Men's 3,500 metres Walk
Athletics	Athletics Men's 4 mile, Team

Medals by country (top)



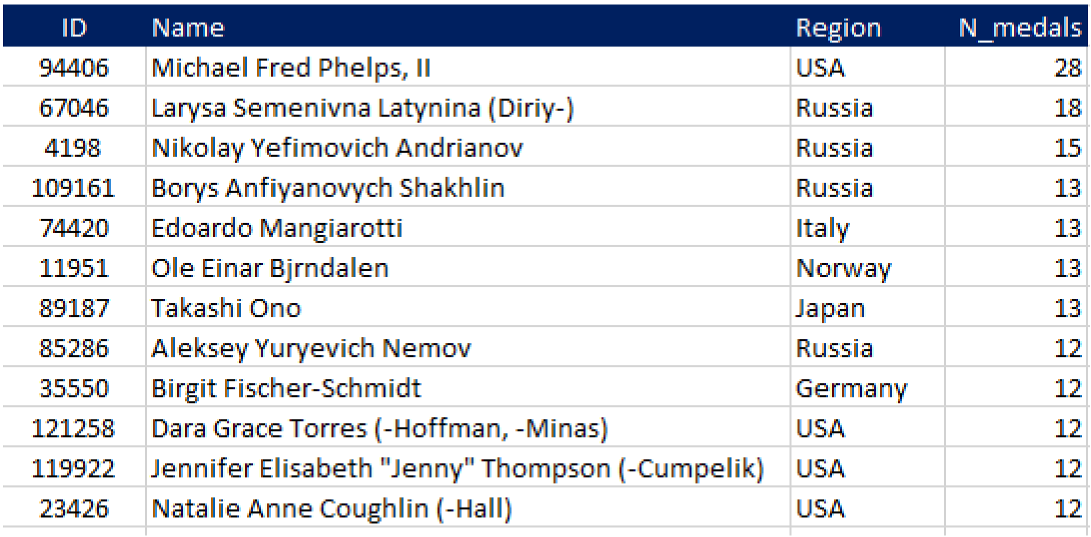
region (tpo 10)	medal	N_medals
USA	Gold	2 638
USA	Silver	1 641
Russia	Gold	1 599
USA	Bronze	1 358
Germany	Gold	1 301
Germany	Bronze	1 260
Germany	Silver	1 195
Russia	Bronze	1 178
Russia	Silver	1 170
UK	Silver	739

Medals in database (85%: NA)



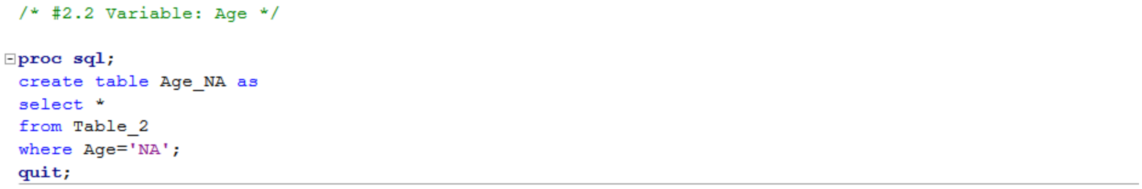
medals	Count_medals	Percent
Bronze	13295	5%
Gold	13369	5%
NA	229959	85%
Silver	13108	5%
	269731	

Top athletes with more medals

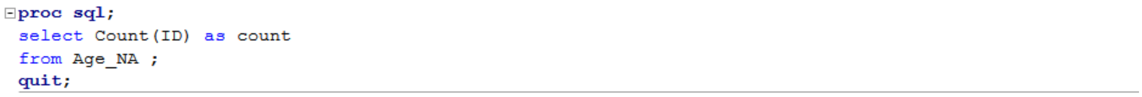


ID	Name	Region	N_medals
94406	Michael Fred Phelps, II	USA	28
67046	Larysa Semenivna Latynina (Diriy-)	Russia	18
4198	Nikolay Yefimovich Andrianov	Russia	15
109161	Borys Anfiyanovych Shakhlin	Russia	13
74420	Edoardo Mangiarotti	Italy	13
11951	Ole Einar Bjrndalen	Norway	13
89187	Takashi Ono	Japan	13
85286	Aleksey Yuryevich Nemov	Russia	12
35550	Birgit Fischer-Schmidt	Germany	12
121258	Dara Grace Torres (-Hoffman, -Minas)	USA	12
119922	Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	USA	12
23426	Natalie Anne Coughlin (-Hall)	USA	12

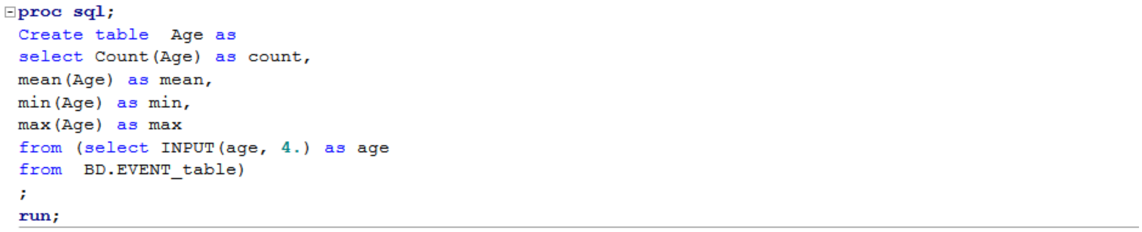
Code (SAS using proc sql) – some examples:



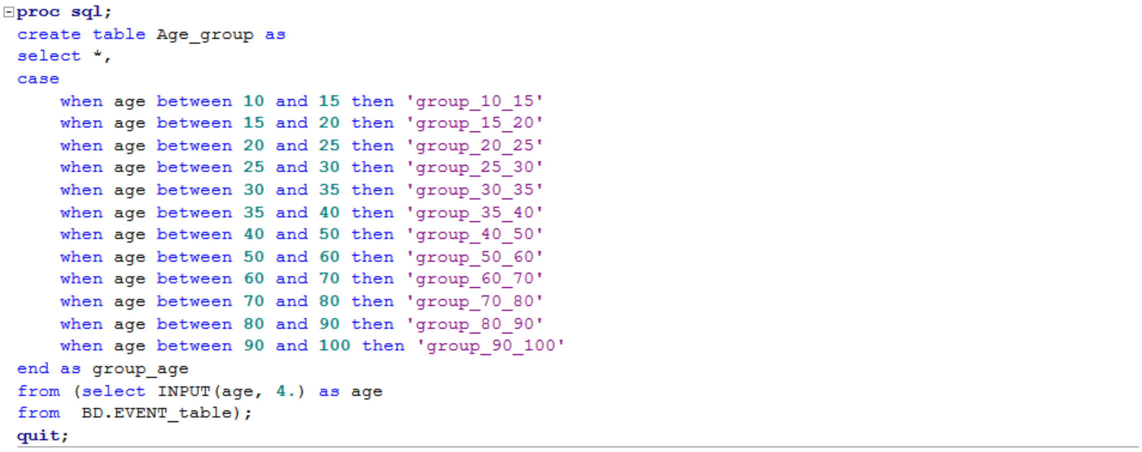
```
/* #2.2 Variable: Age */

proc sql;
create table Age_NA as
select *
from Table_2
where Age='NA';
quit;
```



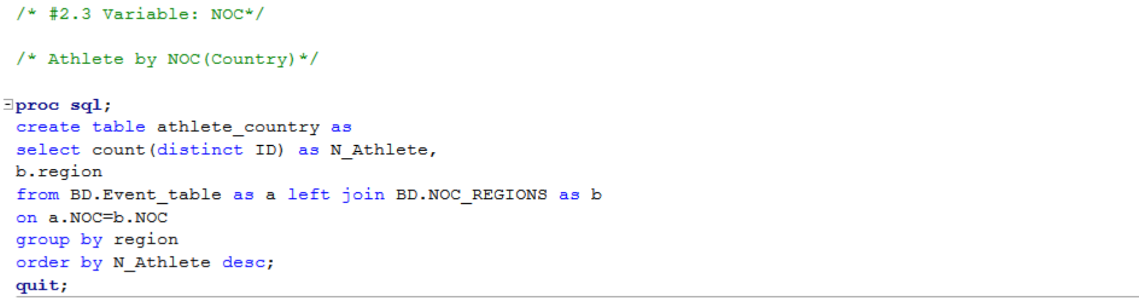
```
proc sql;
select Count(ID) as count
from Age_NA ;
quit;
```



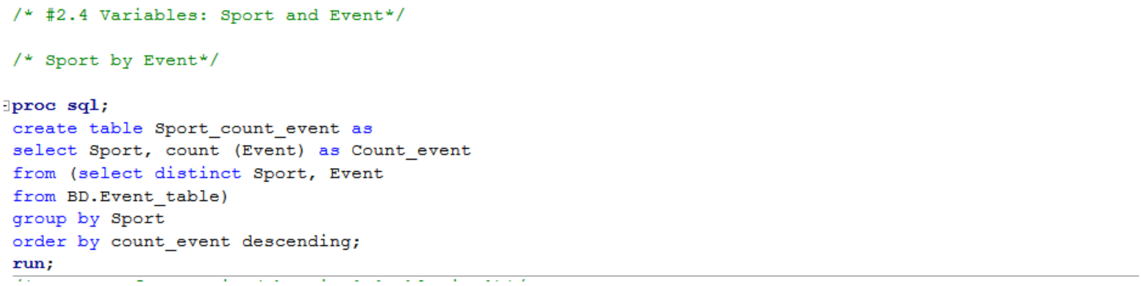
```
proc sql;
Create table  Age as
select Count(Age) as count,
mean(Age) as mean,
min(Age) as min,
max(Age) as max
from (select INPUT(age, 4.) as age
from  BD.EVENT_table)
;
run;
```



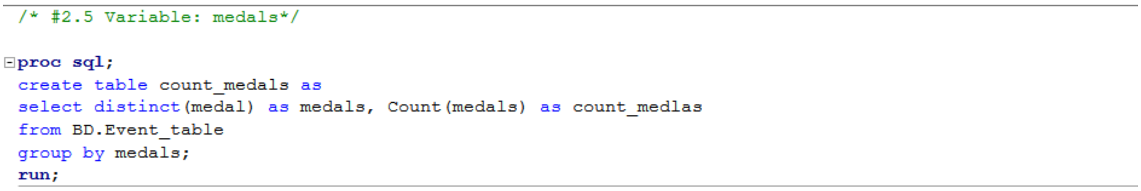
```
proc sql;
create table Age_group as
select *,
case
    when age between 10 and 15 then 'group_10_15'
    when age between 15 and 20 then 'group_15_20'
    when age between 20 and 25 then 'group_20_25'
    when age between 25 and 30 then 'group_25_30'
    when age between 30 and 35 then 'group_30_35'
    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_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 100 then 'group_90_100'
end as group_age
from (select INPUT(age, 4.) as age
from  BD.EVENT_table);
quit;
```



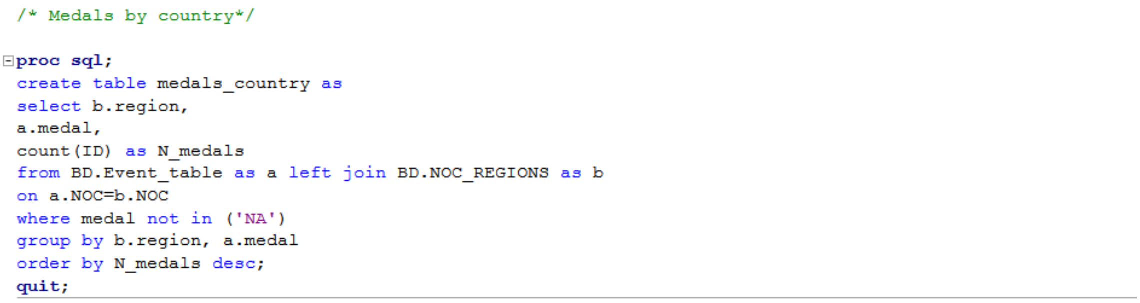
```
/* #2.3 Variable: NOC*/

/* Athlete by NOC(Country)*/

proc sql;
create table athlete_country as
select count(distinct ID) as N_Athlete,
b.region
from BD.Event_table as a left join BD.NOC_REGIONS as b
on a.NOC=b.NOC
group by region
order by N_Athlete desc;
quit;
```



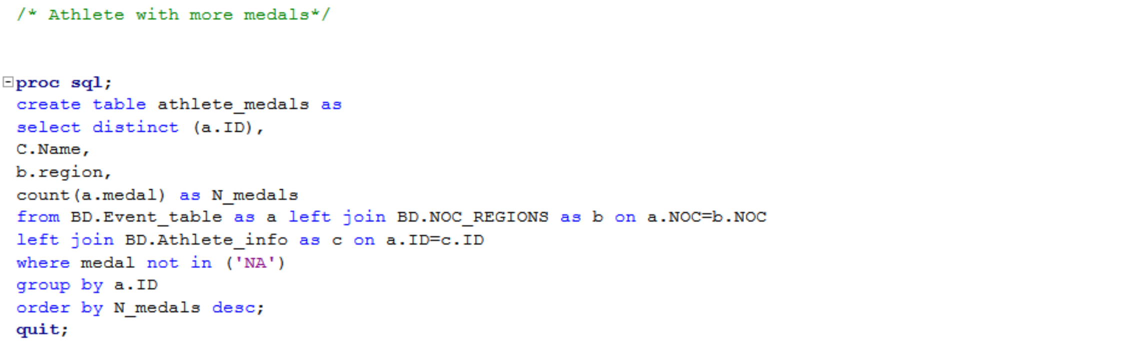
```
/* #2.4 Variables: Sport and Event*/

/* Sport by Event*/

proc sql;
create table Sport_count_event as
select Sport, count (Event) as Count_event
from (select distinct Sport, Event
from BD.Event_table)
group by Sport
order by count_event descending;
run;
```



```
/* #2.5 Variable: medals*/

proc sql;
create table count_medals as
select distinct(medal) as medals, Count(medals) as count_medlas
from BD.Event_table
group by medals;
run;
```



```
/* Medals by country*/

proc sql;
create table medals_country as
select b.region,
a.medal,
count(ID) as N_medals
from BD.Event_table as a left join BD.NOC_REGIONS as b
on a.NOC=b.NOC
where medal not in ('NA')
group by b.region, a.medal
order by N_medals desc;
quit;
```

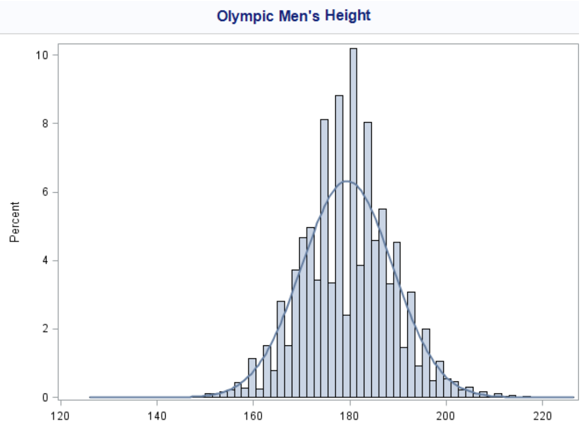


```
/* Athlete with more medals*/


proc sql;
create table athlete_medals as
select distinct (a.ID),
C.Name,
b.region,
count(a.medal) as N_medals
from BD.Event_table as a left join BD.NOC_REGIONS as b on a.NOC=b.NOC
left join BD.Athlete_info as c on a.ID=c.ID
where medal not in ('NA')
group by a.ID
order by N_medals desc;
quit;
```

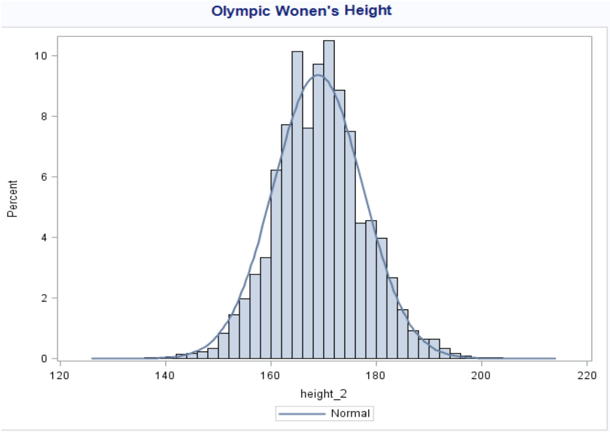
2) Submit 2-3 key points you may have discovered about the data, e.g. new relationships? Did you come up with additional ideas for other things to review?

Some notes:

- Variables gender (sex), height and weight: There is more than twice as many men as women in the database. As expected the average height and weight of males are higher. The maximum and minimum values were surprising, because they are very high and very low, respectively.

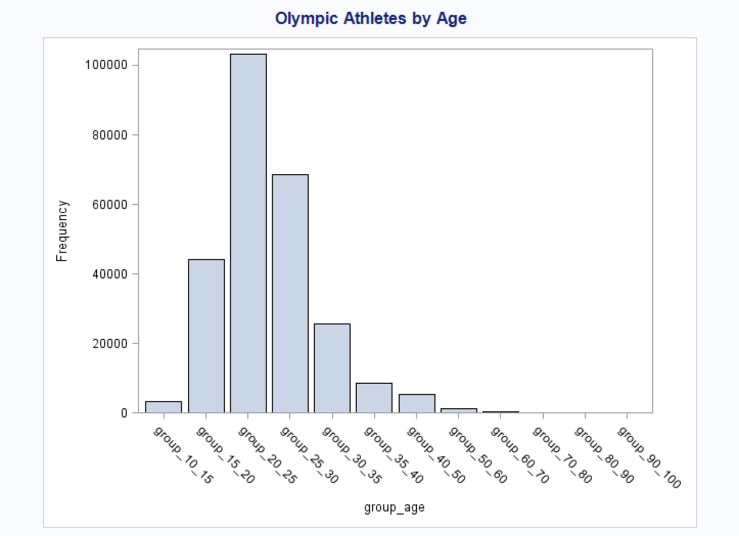


Olympic Men's Height



Olympic Wonen's Height

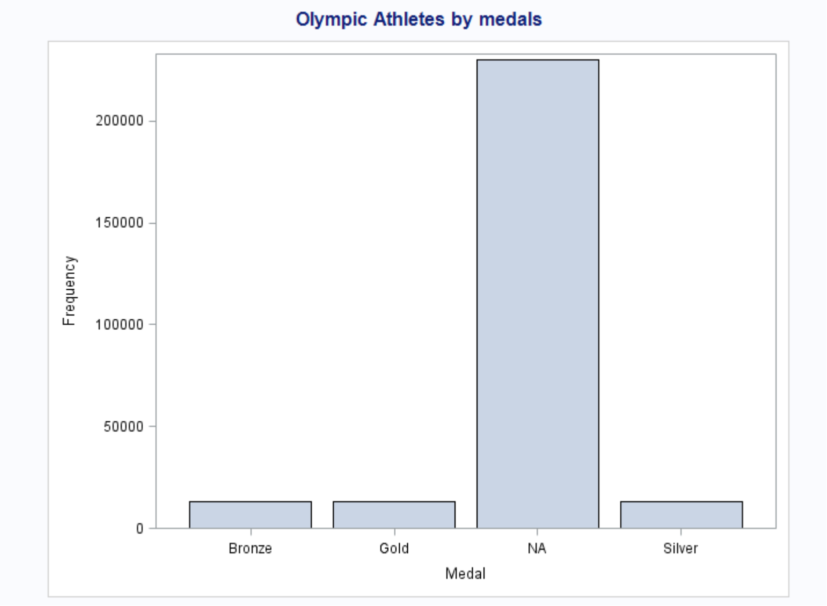
- Variable age: There are very young athletes and others very old. As expected, the vast majority are between 20 and 25 years old.



- Variable country (NOC): The USA is the country with the most participation in events.

- Variable sports and event: In my analysis I will use the Sports instead of the Event. The last variable has a very high level of detail (for example sport Athletics has 83 events: it separates athletics by distances and by gender). However, this level of detail can be useful to test whether the athlete's age can be linked at a distance in athletics events.

- Variable medal: Only 15% of the participants won a medal. The USA, Russia, Germany and the UK are the countries with the most medals won.



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

Hypothesis (week 1)

1) The age group 20-25 is the most represented - confirmed

2) Women in developed countries participate more and get better results (won more medals) – Need more work. Cross sex=f with country (NOC) and medals.

3) US is the most regular country along the years (measure: number of participants and medals) - Partially confirmed. The USA is the country with the most participants and medals (in global terms). However, it is necessary to check over time.

4) Athletics is the sport with more participants. – Need more work. Shooting and Athletics are the sports with the most events, perhaps the ones with the most participants.

5) Russia is the best in gymnastic competitions. Need more work. Cross sport= Gymnastics with country (NOC) and medals.

4) What additional questions are you seeking to answer?

Extra- Questions:

1) Which country (NOC) have the most medals for each sport? Has there been a shift over the time?

2) Are athletes taller today than they were in the past? Is height an advantage in some sports? Is height just as much of an advantage for women as men?