

SQL for Data Science Capstone Project

Assignment 1: WEEK 1 / MILSTONE1 - Project Proposal and Data Selection / Preparation

Name: Daniel Siva

Step 1: Preparing for Your Proposal

1. Which client/dataset did you select and why?

I have chosen the 'SportsStats' with the Olympics dataset, which contains over 120 years of data. I decided to use this dataset because I have some curiosity about this statistics.

2. Describe the steps you took to import and clean the data.

Steps:

- Download the dataset from Coursera
- Extract the CSV from the ZIP file
- Import the data to SAS

SAS code:

```
libname BD "C:\SASstg\DDE\IFM\Data\Controlo_qualidade\Daniel_Silva\Base_Dados_Coursera";  
  
OPTION COMPRESS=YES;  
OPTION VALIDVARNAME=V7;
```

```
proc import datafile="C:\SASstg\DDE\IFM\Data\Controlo_qualidade\Daniel_Silva\Base_Dados_Coursera\athlete_events.csv"  
    out=athlete_events  
    dbms=csv  
    replace;  
    getnames=yes;  
    GUESSINGROWS=5000;  
run;
```

```
proc import datafile="C:\SASstg\DDE\IFM\Data\Controlo_qualidade\Daniel_Silva\Base_Dados_Coursera\noc_regions.csv"  
    out=noc_regions  
    dbms=csv  
    replace;  
    getnames=yes;  
run;
```

- Cleaning the data – eliminate duplicate values and analyzing the tables: ATHLETE_EVENTS and NOC_REGIONS

```
/* #1 - CLEANING THE DATA*/  
  
/* #1.1 Count*/  
  
/* 271.116 obs*/
```

```
proc sql;  
    select Count(ID)  
    from athlete_events;  
quit;
```

```
/*230 obs*/
```

```
proc sql;  
    select Count(Noc)  
    from NOC_REGIONS;  
quit;
```

```
/* # 1.2 - Looking for duplicating rows - DISTINCT BY ALL VARIABLES*/

/* Option: Clean data from duplicates - 269.731 distinct obs (1.385 duplicate obs)*/
```

```
proc sql;
create table athlete_events as
select distinct ID,
Name,
Sex,
AGE,
Height,
Weight,
Team,
NOC,
Games,
Year,
Season,
City,
Sport,
Event,
Medal
from athlete_events;
quit;
```

```
proc sql;
select Count(ID)
from distinct_athlete_events;
quit;
```

- I made different queries in order to understand and correct the information – Main conclusions in the table:

Table: Athlete_Events

	Type	Length	Format	Informat	Notes
ID	Character	7	\$7.	\$7.	
Name	Character	78	\$78.	\$78.	
Sex	Character	3	\$3.	\$3.	
Age					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	2	\$2.	\$2.	
Weight	Character	4	\$4.	\$4.	
Team	Character	6	\$6.	\$6.	
NOC	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)
Games	Character	5	\$5.	\$5.	Same ID (Athlete) could have different NOC (option: do not make sense - maybe error in data)
Year	Character	13	\$13.	\$13.	Concatenation between Year and Season
Season	Numeric	8	BEST12.	BEST32.	
City	Character	8	\$8.	\$8.	
Sport	Character	30	\$30.	\$30.	Detected one error (Games= 1956 Summer, city= Melbourne (not Stockholm)
Event	Character	34	\$34.	\$34.	Same ID could have different sports (ex: water polo, swimming)
Medal	Character	87	\$87.	\$87.	
	Character	10	\$10.	\$10.	

```
/*create the table games_info*/
```

```
proc sql;
create table Games as
select distinct
Games,
Year,
Season,
City
from athlete_events;
quit;
```

```
/* correct a value*/
```

```
proc sql;
update Games
set city='Melbourne'
where Games = '1956 Summer';
quit;
```

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

```
/* Athlete by 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;
```




First results:

	 N_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	Netherlands
14	2883	Switzerland
15	2782	Czech Republic
16	2761	Hungary
17	2637	Spain
18	2393	South Korea
19	2347	Finland
20	2337	Austria
21	2216	Norway
22	2078	Belgium
23	2053	Brazil
24	1923	Denmark
25	1848	Argentina

```
/* 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;
```

First results:

	 region	 Medal	 N_medals
1	USA	Gold	2638
2	USA	Silver	1641
3	Russia	Gold	1599
4	USA	Bronze	1358
5	Germany	Gold	1301
6	Germany	Bronze	1260
7	Germany	Silver	1195
8	Russia	Bronze	1178
9	Russia	Silver	1170
10	UK	Silver	739
11	UK	Gold	677
12	France	Bronze	666
13	UK	Bronze	651
14	France	Silver	602
15	Italy	Gold	575
16	Sweden	Bronze	535
17	Italy	Bronze	531
18	Italy	Silver	531
19	Australia	Bronze	522
20	Sweden	Silver	522
21	France	Gold	499
22	Sweden	Gold	479
23	Canada	Gold	463

```
/* 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;
```

First results:

	ID	Name	region	N_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	109161	Borys Anfiyanovych Shakhlin	Russia	13
5	74420	Edoardo Mangiarotti	Italy	13
6	11951	Ole Einar Bjmdalen	Norway	13
7	89187	Takashi Ono	Japan	13
8	85286	Aleksey Yuryevich Nemov	Russia	12
9	35550	Birgit Fischer-Schmidt	Germany	12
10	121258	Dara Grace Torres (-Hoffman, -Minas)	USA	12
11	119922	Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	USA	12
12	23426	Natalie Anne Coughlin (-Hall)	USA	12
13	87390	Paavo Johannes Nurmi	Finland	12
14	70965	Ryan Steven Lochte	USA	12
15	57998	Sawao Kato	Japan	12
16	89706	Carl Townsend Osburn	USA	11
17	113912	Mark Andrew Spitz	USA	11
18	11642	Matthew Nicholas "Matt" Biondi	USA	11
19	21402	Viktor Ivanovych Chukarin	Russia	11
20	18826	Vra slavsk (-Odloilov)	Czech Republic	11
21	84381	Akinori Nakayama	Japan	10
22	39726	Aladr Gerevich (-Gerei)	Hungary	10

```
/* Games with more observations*/
```

```
proc sql;
create table Games_obs as
select distinct(a.Games), b.City, count(ID) as N_obs
from BD.Event_table as a left join BD.Games_info as b
on a.Games=b.Games
group by b.Games
order by N_obs desc;
quit;
```

First results:

	Games	City	N_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	1948 Summer	London	6308
18	1936 Summer	Berlin	6251
19	1956 Summer	Melbourne	5127
20	1924 Summer	Paris	5110
21	2014 Winter	Sochi	4891
22	1928 Summer	Amsterdam	4656
23	2010 Winter	Vancouver	4402
24	2006 Winter	Torino	4382
25	1920 Summer	Antwerpen	4292
26	2002 Winter	Salt Lake City	4109

/* Athlete by gender*/

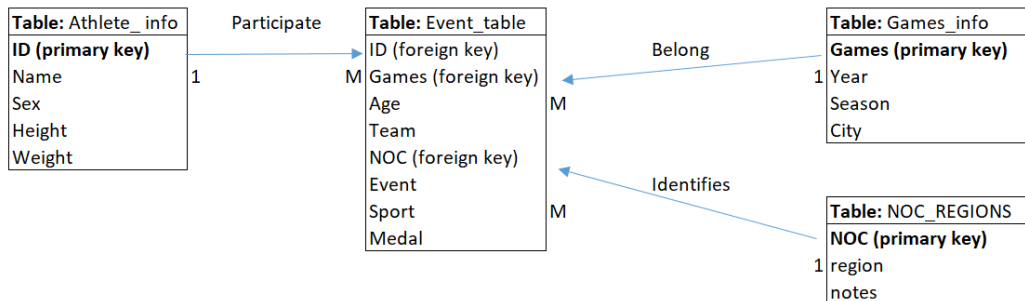
```
proc sql;
  create table gender as
  select sex, count(ID) as N_Athlete
  from BD.Athlete_info
  group by sex;
quit;
```

Results:

	Sex	N_Athlete
1	F	33981
2	M	101590

4. Create an ERD or proposed ERD to show the relationships of the data you are exploring.

Chen Notation



Step 2: Develop Project Proposal

Description:

My project goal is to learn more about Olympic Games results (medals). I hope to find evidences about the country, gender, sex and years to help to understand the results.

Journalists, coaches, countries (government) who are looking to understand the trends in Olympics Games over the years would be interested in my findings in order to publish the information (journalists) or improve results (coaches and countries).

Questions:

- 1) What has been the demographics of Olympics – what age group, sex have more participants and better results?
- 2) Which country has been most regular in Olympics, who has been winning the most medals?
- 3) Which sports in which season Olympics is popular amongst countries?
- 4) Which country has been encouraging female participation in sports over the years? Are there best countries for women participation than men (measure: % of females by country)?
- 5) How has been the dynamics of participation of small countries and which sports are they focusing?

Hypothesis

- 1) The age group 20-25 is the most represented.
- 2) Women in developed countries participate more and get better results (won more medals).
- 3) US is the most regular country along the years (measure: number of participants and medals).
- 4) Athletics is the sport with more participants.
- 5) Russia is better in gymnastic competitions.

Approach

In order to approve/reject the hypothesis I am going to create some bar charts with the tables. I will separate the information in three main data frames, the first one is focused on years, the second in sports and the last one in gender. So, I want to establish a dashboard where the information about the medals and the element of study could be easily seen.