Beyond Descriptive Stats – Olympic Games Dataset

Dive Deeper

Relationships/correlation, Pearson correlation

In the previous assigment I prepared a temporary table that had columns: Year, season, noc, sex, number_participants, medals_number. I could divide my data into four sets: winter-women, winter-men, summer-women and summer-men. I prepared descriptive statistics: average and standard deviation on number_paticipants and medals number. I grouped the data by team. I charted this data in RawGraphs and it seems that there is a correlation between average number of participants and average number of medals. Now it's time to check that.

```
select noc
            ,season
            ,sex
            ,avg(number_participants) as average_part_number
            ,stddev(number_participants) as stdev_part__number
            ,avg(medals_number) as average_medal_number
            ,stddev(medals_number) as stdev_medals_number
from table1  --temporary table I described in text
group by noc,season,sex
--having sex like 'F' and season like 'Summer'  -- uncoment for this data set
--having sex like 'M' and season like 'Summer'
--having sex like 'F' and season like 'Winter'
having sex like 'M' and season like 'Winter' --only one at the time must be
uncoment
order by average_medal_number desc
```

First 20 rows – example for Winter - Men:

noc	season	sex	average_part_number	stdev_part__number	average_medal_number	stdev_medals_number
URS	Winter	M	59	10.8627804912	13.5555555556	4.746343622
RUS	Winter	M	93	18.2098874241	12.5	6.5038450166
NOR	Winter	M	46.1818181818	17.3414899224	11.8636363636	4.6934143785
GDR	Winter	M	37.6666666667	5.3913510984	11	3.8470768123
EUN	Winter	M	86	[NULL]	8	[NULL]
USA	Winter	M	76.8636363636	28.8927333994	7.5	6.6601158291
GER	Winter	M	65.1428571429	22.5656551247	7.5	4.988448194
AUT	Winter	M	48.0454545455	22.0572356614	6.7272727273	4.5688575021
FIN	Winter	M	40.4090909091	15.8674842454	5.4545454545	2.5584085963
SWE	Winter	M	46.9090909091	13.6657691607	4.8181818182	2.5191904149
SUI	Winter	M	47.7727272727	20.9305696193	4.4090909091	2.6305728311

FRG	Winter	M	64.3333333333	5.6450568347	4.3333333333	1.3662601021
CAN	Winter	M	54.3181818182	32.5069756084	4.2272727273	3.9149508077
NED	Winter	M	9.55	5.9247473852	3.5	3.4716900177
ITA	Winter	M	49.6363636364	24.0446914338	3.4090909091	3.0962202904
FRA	Winter	M	42.2272727273	24.088985466	3.2272727273	3.0068464446
JPN	Winter	M	41.1	20.4653751235	1.7	2.364206777
CHN	Winter	M	18.6	9.6976514912	1.7	1.8287822299
CZE	Winter	M	57.6666666667	8.6178110136	1.6666666667	1.3662601021
KOR	Winter	M	14.1176470588	12.2468483341	1.5294117647	2.7412963872

Linear regression for future prediction (if relationship is linear)

Code for computing correlation and regression coefficients:

```
select corr(average_part_number,average_medal_number),
    regr_r2(average_part_number,average_medal_number) AS r2,
    regr_intercept(average_part_number,average_medal_number) AS b,
    regr_slope(average_part_number,average_medal_number) AS a,
    regr_count(average_part_number,average_medal_number) AS cnt
from (
      select noc
                  ,season
                  ,sex
                  ,avg(number_participants) as average_part_number
                  ,stddev(number_participants) as stdev_part__number
                  ,avg(medals_number) as average_medal_number
                  ,stddev(medals_number) as stdev_medals_number
      from table1
      group by noc,season,sex
      having sex like 'M' and season like 'Winter'
--having sex like 'F' and season like 'Summer'
--having sex like 'M' and season like 'Summer'
--having sex like 'F' and season like 'Winter'
) as t0
```

Answers:

Summer – Men

corr	r2	b	a	cnt
0.8885850837	0.789583451	16.789424636	4.5224768144	230

Summer – Women

corr	r2	b	a	cnt
0.87560115984	0.766677391	7.162420414	4.331587744	222

Winter – Men

corr	r2	b	a	cnt
0.82261984471	0.676703409	6.7220461281	6.178770799	114

Winter-Women

corr	r2	b	a	cnt
0.81291837363	0.6608362822	4.2185210832	3.0377191215	90

Now I see that I was right. Average number of participants and average number of medals are strongly correlated, as indicated by large R^2 value.

The other example of linear regression is that percent of women participants is corellated to the Year variable (it's growing over time)

I prepared table of participants

id	Name	sex	Year	season	noc
1	A Dijiang	M	1,992	Summer	CHN
2	A Lamusi	M	2,012	Summer	CHN
3	Gunnar Nielsen Aaby	M	1,920	Summer	DEN
4	Edgar Lindenau Aabye	M	1,900	Summer	DEN
5	Christine Jacoba Aaftink	F	1,988	Winter	NED

I used distinct selection, so one participant could appear in this table several times but not in the same games.

Based on that table I prepared another one to calculate how many percent of participants were women:

```
select "Year"
	,season
	,(100.0*number_of_participants_f)/(number_of_participants_f +
number_of_participants_m) as female_perc
from
	(select m."Year"
		,m.season
		,m.number_of_participants as number_of_participants_m
		,coalesce (f.number_of_participants,0) as number_of_participants_f
		from
			(select "Year"
			,season
			,count(id) as number_of_participants
		from participants p2
		where sex like 'M'
	group by "Year",season,sex) as m
	left join
				(select "Year"
			,season
			,count(id) as number_of_participants
		from participants p2
		where sex like 'F'
	group by "Year",season,sex) as f
```

```
        on m."Year" = f."Year" and m.season=f.season) as tab
```

Year	season	female_perc
1,896	Summer	0
1,900	Summer	1.8790849673
1,904	Summer	0.9230769231
1,906	Summer	0.7134363853
1,908	Summer	2.1739130435
1,912	Summer	2.200083022
1,920	Summer	2.9147982063
1,924	Summer	4.7911547912
1,924	Winter	4.1533546326
1,928	Summer	9.6088697259
1,928	Winter	6.0737527115
1,932	Summer	10.4578563996
1,932	Winter	8.3333333333
1,936	Summer	8.0508474576
1,936	Winter	11.9760479042
1,948	Summer	10.1317582917
1,948	Winter	11.5269461078
1,952	Summer	10.5636658556
1,952	Winter	15.7060518732
1,956	Summer	11.4729608605
1,956	Winter	16.077953715
1,960	Summer	11.4536621824
1,960	Winter	21.6541353383
1,964	Summer	13.2372980339
1,964	Winter	18.281535649
1,968	Summer	14.0878013674
1,968	Winter	18.1896551724
1,972	Summer	14.9001967951
1,972	Winter	20.4365079365
1,976	Summer	20.7475712169
1,976	Winter	20.4787234043
1,980	Summer	21.3538695569
1,980	Winter	21.9421101774
1,984	Summer	23.0803177405
1,984	Winter	21.7596229379
1,988	Summer	26.0586704519
1,988	Winter	22.1052631579
1,992	Summer	29.0112934157
1,992	Winter	27.0960577457
1,994	Winter	30.0345224396
1,996	Summer	34.0168294806

1,998	Winter	36.2092703075
2000	Summer	38.2079459003
2,002	Winter	36.9320550229
2,004	Summer	40.7312683528
2,006	Winter	38.2919005613
2,008	Summer	42.2882833287
2,010	Winter	40.7334384858
2,012	Summer	44.2521631644
2,014	Winter	40.14571949
2,016	Summer	45.0308614366

At Olympic Games in 1896 there were no women participants and now almost half of all participants are female.

```
select corr("Year",female_perc),
    regr_r2("Year",female_perc) AS r2,
    regr_intercept("Year",female_perc) AS b,
    regr_slope("Year",female_perc) AS a,
    regr_count("Year",female_perc) AS cnt
from
      (
            select "Year"
            ,season
            ,(100.0*number_of_participants_f)/(number_of_participants_f +
number_of_participants_m) as female_perc
      from
            (select m."Year"
                  ,m.season
                  ,m.number_of_participants as number_of_participants_m
                  ,coalesce (f.number_of_participants,0) as
number_of_participants_f
                  from
                        (select "Year"
                        ,season
                        ,count(id) as number_of_participants
                  from participants p2
                  where sex like 'M'
            group by "Year",season,sex) as m
            left join
                              (select "Year"
                        ,season
                        ,count(id) as number_of_participants
                  from participants p2
                  where sex like 'F'
            group by "Year",season,sex) as f
            on m."Year" = f."Year" and m.season=f.season) as tab
      ) as tab1
```

Corr	r2	B	a	cnt
0.9520753909	0.90644755	1,915.8001528124	2.4633563774	51

Percent of women participants is strongly correlated to Year.

Textual Analysis for TF-IDF (Term Frequency-Inverse Document Frequency; Row-based and column-based, stop-word removal?

I didn't expect that I would need to prepare TF-ID analysis for Olympic Games dataset.

I will prepare this using column „Event".

At first I prepared tables:

```
create temporary table table2
 as
	select distinct "Event",sport,"Year", season from athlete_events_csv;
```

Before I prepared next table, I checked that every row has one of the words: Mixed,Women's or Men's. Checking word freqency for every year would be laborious and I believe that I don't need so specific information, so I decided to add „Year_range" and check the word frequency for larger periods.

```
create temporary table table3
as
	select *
		,case when"Event" like '%Mixed%' then 'mixed'
		when "Event" like '%Women%' then 'F'
		else  'M'
		end as gender
		,case when "Year" <= 1916 then '1896-1916'
		when "Year">1916 and "Year"<= 1936 then '1916-1936'
		when "Year">1936 and "Year"<= 1956 then '1936-1956'
		when "Year">1956 and "Year"<= 1976 then '1956-1976'
		when "Year">1976 and "Year"<= 1996 then '1976-1996'
		else '1996-2016'
		end as "Year_range"
	from table2
```

First 20 rows:

Event	sport	Year	season	gender	Year_range
Diving Men's Springboard	Diving	1,984	Summer	M	1976-1996
Equestrianism Mixed Jumping, Team	Equestrianism	1,956	Summer	mixed	1936-1956
Athletics Men's 10,000 metres	Athletics	1,92	Summer	M	1916-1936
Cross Country Skiing Women's 10 kilometres	Cross Country Skiing	2,014	Winter	F	1996-2016
Canoeing Women's Kayak Singles, Slalom	Canoeing	1,992	Summer	F	1976-1996
Biathlon Women's 4 x 7.5 kilometres Relay	Biathlon	1,998	Winter	F	1996-2016
Wrestling Men's Middleweight, Greco-Roman	Wrestling	1,992	Summer	M	1976-1996
Boxing Men's Super-Heavyweight	Boxing	2,004	Summer	M	1996-2016
Athletics Women's Long Jump	Athletics	1,98	Summer	F	1976-1996
Wrestling Men's Bantamweight, Greco-Roman	Wrestling	1,932	Summer	M	1916-1936
Football Men's Football	Football	1,906	Summer	M	1896-1916

Figure Skating Women's Singles	Figure Skating	1,924	Winter	F	1916-1936
Swimming Women's 4 x 100 metres Medley Relay	Swimming	1,964	Summer	F	1956-1976
Boxing Men's Light-Heavyweight	Boxing	1,98	Summer	M	1976-1996
Swimming Women's 100 metres Freestyle	Swimming	1,932	Summer	F	1916-1936
Cycling Men's BMX	Cycling	2,016	Summer	M	1996-2016
Judo Women's Half-Lightweight	Judo	1,996	Summer	F	1976-1996
Athletics Women's 4 x 100 metres Relay	Athletics	1,98	Summer	F	1976-1996
Cycling Women's Points Race	Cycling	1,996	Summer	F	1976-1996
Swimming Men's 4 x 200 metres Freestyle Relay	Swimming	1,956	Summer	M	1936-1956

For this table I there are almost no stop words but there are still some words that I would like to remove.

At first I tried to use replace function:

```
select "Event"
	,replace(replace(replace(replace(replace(replace(replace(replace(replace
("Event",' Mixed ',' '),' kilometres ',' '),' kilometres',''),' Men''s ',' '),'
Women''s ',' '),' x ',' '), ' 4 ',' '),' metres ',' '),' metres','')
	,sport
	,"Year"
	,season
	,gender
	,"Year_range"
from table3
```

But this code was too long and I didn't even put there all words that I wanted to remove.

Without resorting to Python it's easier to split words first and remove the unwanted ones later:

```
select *
from
	(select "Year_range"
		,gender
		,season
		,"Event"
		,unnest(string_to_array("Event",' ')) as text
	from table3) as tab
where text not in
('Mixed','kilometres','Men''s','Women''s','metres','x','10,000','4','10','7.5','Si
ngles','Pairs','Two','Three','Person''Singles,','Team','100','200','300','1,000','
50','500','20','20+','25','18','15','1,980','yards','Fours','Free','1,500','3,000'
)
```

Now I am ready to count words.

```
select text
	,count(text) as n
from
	(select *
```

```
	from
		(select "Year_range"
			,gender
			,season
			,"Event"
			,unnest(string_to_array("Event",' ')) as text
		from table3) as tab1
	where text not
in('Mixed','kilometres','Men''s','Women''s','metres','x','10,000','4','10','7.5','
Singles','Pairs','Two','Three','Person''Singles,','Team','100','200','300','400','
1,000','50','500','20','20+','25','18','15','1,980','yards','Fours','Free','1,500'
,'3,000','Indyvidual')) as tab2
group by text
order by n desc
limit(20)
```

text	n
Athletics	998
Swimming	580
Freestyle	514
Individual	508
Wrestling	414
Gymnastics	336
Skiing	333
Skating	307
Shooting	285
Rowing	261
Relay	253
Boxing	252
Cycling	248
Canoeing	226
Speed	222
Fencing	219
Weightlifting	215
Greco-Roman	198
Sailing	173
Cross	161

I expect that most frequented words will be different for summer and for winter games. I Wonder if there is any difference regarding gender.

```
select text
	,count(text) as n
from
	(select *
	from
		(select "Year_range"
			,gender
			,season
			,"Event"
```

```
                    ,unnest(string_to_array("Event",' ')) as text
               from table3) as tab1
        where text not
in('And','Mixed','kilometres','Men''s','Women''s','metres','x','10,000','4','10','
7.5','5','Singles','Pairs','Doubles','One','Two','Three','Person','Singles,','Team
','100','200','300','400','1,000','5,000','50','500','20','20+','25','18','15','1,
980','yards','Fours','Free','1,500','3,000','Indyvidual')) as tab2
--where gender like 'F' and season like 'Summer'
--where gender like 'M' and season like 'Summer'
--where gender like 'F' and season like 'Winter'
--where gender like 'M' and season like 'Winter'
--where gender like 'mixed' and season like 'Summer'
--where gender like 'mixed' and season like 'Winters' -- no such case
group by text
order by n desc
limit(20)
```

Summer - women

text	n
Athletics	307
Swimming	270
Freestyle	129
Individual	127
Gymnastics	122
Relay	78
Rowing	66
Diving	57
Canoeing	52
Kayak	52
Fencing	52
Cycling	51
Judo	49
Shooting	48
Throw	46
Jump	45
Tennis	42
Medley	40
Volleyball	40
Sculls	39

Summer-men

text	n
Athletics	691
Wrestling	396
Freestyle	351
Swimming	310
Individual	272
Boxing	246
Gymnastics	214
Shooting	202
Greco-Roman	198
Cycling	197
Rowing	195
Weightlifting	180
Canoeing	174
Fencing	167
Relay	104
Jump	99
Kayak	93
Judo	88
Throw	84
Rifle,	84

Summer-mixed

text	n
Sailing	117
Equestrianism	98
Competitions	78
Art	78
Individual	48
Shooting	35
Jumping,	33
Dressage,	33
Dinghy	32
Keelboat	30
Event,	28
Three-Day	28
Painting,	16
Unknown	13
Sculpturing,	13
Event	13
Architecture,	11
Literature,	10
Rifle,	10
Heavyweight	10

Winter-women

text	n
Skiing	155
Skating	117

Winter-men

text	n
Skiing	178
Skating	149

Speed	95
Alpine	71
Cross	69
Country	67
Slalom	41
Relay	30
Biathlon	28
Short	24
Track	24
Figure	22
Giant	22
Downhill	18
Freestyle	17
Snowboarding	15
Luge	14
Sprint	14
Pursuit	11
Combined	10

Speed	127
Cross	92
Country	90
Alpine	71
Individual	60
Ski	48
Biathlon	45
Combined	44
Hill,	44
Jumping	44
Ice	44
Hockey	44
Slalom	41
Bobsleigh	40
Relay	39
Nordic	34
Short	24
Normal	24

I can see that there are differences between summer and winter games but not in women-men events. In the next step I will check separately summer and winter words for prepared periods of time.

```
select text
	,count(text) as n
from
	(select *
	from
		(select "Year_range"
			,gender
			,season
			,"Event"
			,unnest(string_to_array("Event",' ')) as text
		from table3) as tab1
	where text not
in('And','Mixed','kilometres','Men''s','Women''s','metres','x','10,000','4','10','
7.5','5','Singles','Pairs','Doubles','One','Two','Three''Four','Four/Five','Person
','Singles,','Team','100','200','300','400','1,000','5,000','50','500','20','20+',
'25','18','15','1,980','yards','Fours','Free','1,500','3,000','Indyvidual')) as
tab2
--where "Year_range" like '1896-1916' and season like 'Summer'
--where "Year_range" like '1916-1936' and season like 'Summer'
--where "Year_range" like '1936-1956' and season like 'Summer'
--where "Year_range" like '1956-1976' and season like 'Summer'
--where "Year_range" like '1976-1996' and season like 'Summer'
--where "Year_range" like '1996-2016' and season like 'Summer'
--where "Year_range" like '1896-1916' and season like 'Winter'  --no such case
first 1924
--where "Year_range" like '1916-1936' and season like 'Winter'
--where "Year_range" like '1936-1956' and season like 'Winter'
--where "Year_range" like '1956-1976' and season like 'Winter'
```

```
--where "Year_range" like '1976-1996' and season like 'Winter'
where "Year_range" like '1996-2016' and season like 'Winter'
group by text
order by n desc
limit(20)
```

Summer

1896-1916

text	n
Athletics	136
Shooting	58
Individual	45
Swimming	42
Freestyle	42
Jump	32
Fencing	32
Gymnastics	31
Cycling	30
Rifle,	28
Wrestling	26
Tennis	26
Rowing	25
Archery	17
Sailing	16
High	14
Greco-Roman	14
Throw	14
Pistol,	13
Standing	12

1916-1936

text	n
Athletics	141
Freestyle	68
Wrestling	64
Individual	57
Art	55
Competitions	55
Swimming	54
Gymnastics	41
Boxing	40
Shooting	37
Fencing	34
Rowing	33
Greco-Roman	31
Cycling	30
Equestrianism	30
Weightlifting	25
Sailing	24
Relay	23
Diving	22
Throw	21

1936-1956

text	n
Athletics	99
Wrestling	48
Freestyle	45
Gymnastics	39
Swimming	35
Individual	32
Boxing	28
Canoeing	27
Greco-Roman	24
Rowing	21
Fencing	21
Weightlifting	20
Cycling	18
Shooting	18
Equestrianism	18
Competitions	18
Art	18
Throw	15
Jump	15
Relay	15

1956-1976

text	n
Athletics	181
Swimming	117
Freestyle	91
Wrestling	88
Individual	75
Gymnastics	70
Boxing	53
Greco-Roman	44
Canoeing	43
Rowing	42

1976-1996

text	n
Athletics	208
Swimming	156
Freestyle	111
Individual	101
Wrestling	100
Gymnastics	75
Rowing	70
Canoeing	67
Shooting	59
Boxing	59

1996-2016

text	n
Athletics	233
Swimming	176
Individual	137
Freestyle	123
Cycling	90
Wrestling	88
Canoeing	80
Gymnastics	80
Shooting	79
Weightlifting	75

Relay	41
Fencing	40
Weightlifting	39
Shooting	34
Cycling	33
Equestrianism	29
Kayak	29
Sailing	25
Throw	25
Jump	25

Judo	51
Greco-Roman	50
Weightlifting	50
Cycling	47
Relay	45
Kayak	43
Fencing	42
Sailing	41
Sculls	32
Equestrianism	30

Judo	70
Rowing	70
Boxing	60
Sailing	53
Kayak	52
Fencing	50
Relay	50
Lightweight	42
Tennis	42
Sculls	40

Winter

1916-1936

text	n
Skating	29
Speed	17
Figure	12
Skiing	11
Country	9
Cross	9
Individual	8
Ice	8
Hockey	8
Ski	6
Combined	6
Bobsleigh	6
Hill,	4
Nordic	4
Normal	4
Jumping	4
Patrol	2
Skeleton	2
Alpine	2
Military	2

1936-1956

text	n
Skiing	31
Skating	21
Alpine	18
Country	13
Cross	13
Speed	12
Slalom	10
Figure	9
Downhill	6
Hockey	6
Individual	6
Ice	6
Bobsleigh	6
Combined	5
Relay	4
Giant	4
Hill,	3
Normal	3
Ski	3
Nordic	3

1956-1976

text	n
Skiing	64

1976-1996

text	n
Skiing	91

1996-2016

text	n
Skiing	136

Skating	57
Speed	41
Cross	34
Country	34
Alpine	30
Slalom	20
Figure	16
Individual	14
Relay	13
Luge	12
Ice	11
Downhill	10
Giant	10
Hockey	10
Hill,	9
Ski	9
Jumping	9
Biathlon	8
Bobsleigh	8

Skating	78
Speed	58
Cross	43
Country	43
Alpine	42
Biathlon	21
Relay	21
Slalom	20
Figure	20
Individual	15
Luge	15
Ice	15
Combined	14
Ski	13
Hill,	13
Jumping	13
Giant	10
Bobsleigh	10
Hockey	10

Skating	115
Speed	94
Cross	62
Country	58
Alpine	50
Biathlon	45
Track	38
Short	38
Relay	32
Slalom	32
Snowboarding	30
Freestyle	28
Sprint	26
Ice	25
Combined	24
Figure	21
Giant	20
Ski	20
Curling	20

Now I can see which sports had most sport events in the olympic games in given periods. I see that there are some changes but not much of them.

Text analysis shows which sports were present on Games during the time. Which sports disciplines have more events for women and which for men. The problem is that with the data I have it's easier and more accurate to do this by analysing column sport.

Go Broader

When I was preparing this analysis I thought about number of participants in given year. Now I can think that it would be good to check if the number of participanting teams is increasing over years. Another question is which of the team has the highest percent of women participants. Are there any that don't have women team?

```
select "Year"
      ,season
      ,count(distinct noc) as number_of_teams
from participants
group by "Year",season
```

Year	season	number_of_teams
1,896	Summer	12
1,900	Summer	31

1,904	Summer	15
1,906	Summer	21
1,908	Summer	22
1,912	Summer	29
1,920	Summer	29
1,924	Summer	45
1,924	Winter	19
1,928	Summer	46
1,928	Winter	25
1,932	Summer	47
1,932	Winter	17
1,936	Summer	49
1,936	Winter	28
1,948	Summer	59
1,948	Winter	28
1,952	Summer	69
1,952	Winter	30
1,956	Summer	72
1,956	Winter	32
1,960	Summer	84
1,960	Winter	30
1,964	Summer	93
1,964	Winter	36
1,968	Summer	112
1,968	Winter	37
1,972	Summer	121
1,972	Winter	35
1,976	Summer	92
1,976	Winter	37
1,980	Summer	80
1,980	Winter	37
1,984	Summer	140
1,984	Winter	49
1,988	Summer	159
1,988	Winter	57
1,992	Summer	169
1,992	Winter	64
1,994	Winter	67
1,996	Summer	197
1,998	Winter	72
2000	Summer	200
2,002	Winter	77
2,004	Summer	201
2,006	Winter	79
2,008	Summer	204

2,010	Winter	82
2,012	Summer	205
2,014	Winter	89
2,016	Summer	207

The answer is yes, number of teams is growing over time.

```
select        avg(number_of_teams)
              ,min(number_of_teams)
              ,max(number_of_teams)
from (
       select "Year"
              ,season
              ,count(distinct noc) as number_of_teams
       from participants
       group by "Year",season ) as tab
where season like 'Summer'
--where season like 'Winter'
```

Summer

avg	min	max
96.8965517241	12	207

Winter

avg	min	max
46.6818181818	17	89

```
select noc
       ,season
       ,(100.0*number_of_participants_f)/(number_of_participants_f +
number_of_participants_m) as female_perc
from
       (select m.noc
              ,m.season
              ,m.number_of_participants as number_of_participants_m
              ,coalesce (f.number_of_participants,0) as number_of_participants_f
              from
                     (select noc
                     ,season
                     ,count(id) as number_of_participants
              from participants p2
              where sex like 'M'
       group by noc,season,sex) as m
       left join
                           (select noc
                     ,season
                     ,count(id) as number_of_participants
              from participants p2
              where sex like 'F'
       group by noc,season,sex) as f
```

```
        on m.noc = f.noc and m.season=f.season) as tab
```

First 20 rows:

noc	season	female_perc
AFG	Summer	4.132231405
AHO	Summer	15.3846153846
AHO	Winter	0
ALB	Summer	37.5
ALB	Winter	33.3333333333
ALG	Summer	17.3553719008
ALG	Winter	28.5714285714
AND	Summer	25
AND	Winter	23.9130434783
ANG	Summer	48.347107438
ANT	Summer	26.6666666667
ANZ	Summer	3.6363636364
ARG	Summer	17.4464363795
ARG	Winter	23.2954545455
ARM	Summer	12.2580645161
ARM	Winter	35.4838709677
ARU	Summer	34.2105263158
ASA	Summer	20
ASA	Winter	0
AUS	Summer	33.8329764454

Now I will check descriptive stats:

```
select avg(female_perc)
       ,min(female_perc)
       ,max(female_perc)
from (
       select noc
              ,season
              ,(100.0*number_of_participants_f)/(number_of_participants_f +
number_of_participants_m) as female_perc
       from
              (select m.noc
                     ,m.season
                     ,m.number_of_participants as number_of_participants_m
                     ,coalesce (f.number_of_participants,0) as
number_of_participants_f
                     from
                            (select noc
                            ,season
                            ,count(id) as number_of_participants
                     from participants p2
                     where sex like 'M'
```

```
            group by noc,season,sex) as m
            left join
                              (select noc
                        ,season
                        ,count(id) as number_of_participants
                  from participants p2
                  where sex like 'F'
            group by noc,season,sex) as f
            on m.noc = f.noc and m.season=f.season) as tab
            ) as tab1
where season like 'Summer'
--where season like 'Winter'
```

Summer

avg	min	max
24.0928672197	0	62.5

Winter

avg	min	max
20.5519064359	0	80

There are many teams that never had women participants but there are also teams where most of the participants were women.

New Metric

What about place in games and rank with number of participants? Those two metrics, especially compared together, can shows us if the country with highest number of participants will have highest number of medals.

```
select rank() over (partition by g2."Year" order by count(g2.participant_id) desc)
as number_participants_rank
      ,g2."Year"
      ,g2.season
      ,p2.noc
      ,count(g2.participant_id) as number_participants
from games g2
left join participants p2
on g2.participant_id = p2.id and g2."Year" = p2."Year" and g2.season = p2.season
group by g2."Year", p2.noc, g2.season
order by g2."Year", number_participants desc;
```

number_participants_rank	Year	season	noc	number_participants
1	1,896	Summer	GRE	102
2	1,896	Summer	GER	19

3	1,896	Summer	USA	14
4	1,896	Summer	FRA	12
5	1,896	Summer	GBR	10
6	1,896	Summer	HUN	7
7	1,896	Summer	SUI	3
7	1,896	Summer	DEN	3
7	1,896	Summer	AUT	3
10	1,896	Summer	SWE	1
10	1,896	Summer	AUS	1
10	1,896	Summer	ITA	1
1	1,900	Summer	FRA	720
2	1,900	Summer	GBR	104
3	1,900	Summer	GER	76
4	1,900	Summer	USA	75
5	1,900	Summer	BEL	64
6	1,900	Summer	NED	35
7	1,900	Summer	ITA	23
8	1,900	Summer	HUN	18

And for medals:

```
select rank() over (partition by "Year" order by count(medal) desc) as game_rank
            ,"Year"
            ,season
            ,noc
            ,count(medal) as medals_number
from medals m2
group by "Year",season, noc
order by "Year", medals_number desc;
```

game_rank	Year	season	noc	medals_number
1	1,896	Summer	GRE	44
2	1,896	Summer	USA	19
3	1,896	Summer	GER	14
4	1,896	Summer	FRA	11
5	1,896	Summer	GBR	9
6	1,896	Summer	HUN	6
6	1,896	Summer	DEN	6
8	1,896	Summer	AUT	5
9	1,896	Summer	SUI	3
9	1,896	Summer	AUS	3
1	1,900	Summer	FRA	102
2	1,900	Summer	USA	54
3	1,900	Summer	GBR	43
4	1,900	Summer	BEL	18
5	1,900	Summer	SUI	11
6	1,900	Summer	GER	8

7	1,900	Summer	DEN	7
7	1,900	Summer	NED	7
9	1,900	Summer	AUT	6
9	1,900	Summer	AUS	6

In the table above I don't have teams with the lowest rank, I mean when the team didn't get any medal at all. It's easy to show them in joined table:

```
select pr."Year"
       ,pr.season
       ,pr.noc
       ,pr.number_participants
       ,coalesce(gr.medals_number,0) as medals_number
       ,pr.number_participants_rank
       ,coalesce(gr.game_rank,max(number_participants_rank) over (partition by
pr."Year",pr.season))as game_rank
       ,(pr.number_participants_rank -
coalesce(gr.game_rank,max(number_participants_rank) over (partition by
pr."Year",pr.season))) as participants_medals_diff
from (select rank() over (partition by g2."Year",g2.season order by
count(g2.participant_id) desc) as number_participants_rank
                    ,g2."Year"
                    ,g2.season
                    ,p2.noc
             ,count(g2.participant_id) as number_participants
             from games g2
             left join participants p2
             on g2.participant_id = p2.id and g2."Year" = p2."Year" and g2.season
= p2.season
             group by g2."Year", p2.noc,g2.season
             order by g2."Year", number_participants desc) as pr
left join (select rank() over (partition by "Year",season order by count(medal)
desc) as game_rank
             ,"Year"
             ,season
             ,noc
             ,count(medal) as medals_number
             from medals m2
       group by "Year", noc, season
       order by "Year", medals_number desc) as gr
on pr."Year" = gr."Year" and pr.season =gr.season and pr.noc = gr.noc
```

Year	season	noc	number_participants	medals_number	number_participants_rank	game_rank	participants_medals_diff
1,896	Summer	SWE	1	0	10	10	0
1,896	Summer	USA	14	19	3	2	1
1,896	Summer	GRE	102	44	1	1	0
1,896	Summer	GER	19	14	2	3	-1
1,896	Summer	AUS	1	3	10	9	1
1,896	Summer	AUT	3	5	7	8	-1
1,896	Summer	HUN	7	6	6	6	0
1,896	Summer	GBR	10	9	5	5	0
1,896	Summer	ITA	1	0	10	10	0
1,896	Summer	FRA	12	11	4	4	0

1,896	Summer	DEN	3	6	7	6	1
1,896	Summer	SUI	3	3	7	9	-2
1,900	Summer	BEL	64	18	5	4	1
1,900	Summer	ESP	9	1	13	19	-6
1,900	Summer	IRI	1	0	22	22	0
1,900	Summer	NZL	1	1	22	19	3
1,900	Summer	BRA	1	0	22	22	0
1,900	Summer	NOR	7	5	14	11	3
1,900	Summer	ARG	1	0	22	22	0
1,900	Summer	LUX	1	1	22	19	3