Sending more participants gives more medals - analysis of SportsStats (Olympic Dataset – 120 years of data)

Data Analysis Project for SQL for Data Science Capstone Project Course

March 2021

Anna Witkowiak

Client

This presentation is planned to be seen by my colleagues from this course. I believe that they will be interested in my approach to analyse Olympic Games Data Set the same way I’m interested in their analysies.

Questions to Answer:

1. How many participants are sent by countries?

- Number of participants can change over time.
- Some countries can send more participants than the others

2. How many medals for country in a specific year?

- This will show if the countries with higher rank in number of participants will have higher rank in number in medals

Initial Hypotheses

1. Number of participants will be larger every year.

- Traveling is less expensive and easier every year so countries can send more participants.

2. The same teams will have highest ranks year after year. Those will be the same countries that send more participants than the others.

- If you can afford to send more participants, you can afford to prepare them in the proper way.

3. There will be more men than women participants but this will change over time

- For many countries men’s medals are more important than women’s

Data Analysis Approach

1. I will add ranks for the countries by number of participants and by number of medals.

- I will focus on descriptive stats of difference between ranks

Data Analysis Approach

2. I will build a table with number of medals and number of participants. This table will also consist information about year, season, olympic team (noc) and sex.

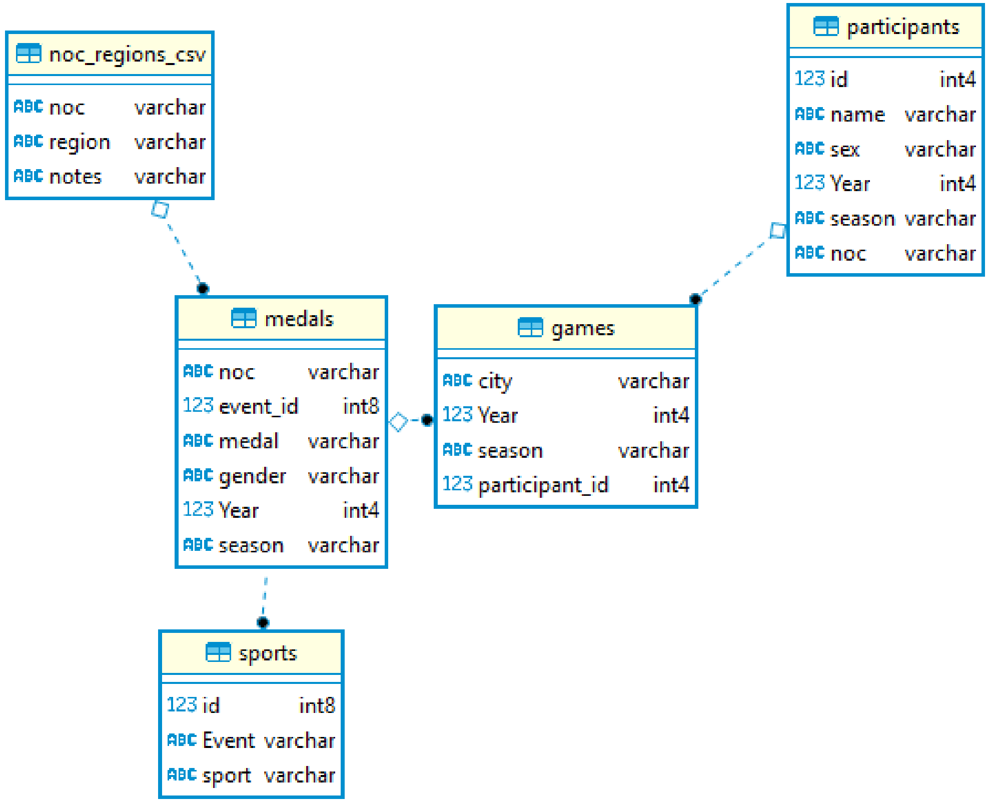
- I will check average and standard deviation for number of medals and number of participants.
- I will divide my set into four groups: Summer-Men, Summer-Women, Winter-Men, Winter-Women.
- I will check correlation between average number of participants and average number of medals

Data Analysis Approach

3. I will check how percent of women participants will change over time.

- I will check how many percent of women participants were in every game (year and season)
- I will also check this information by country – how many percent of women participants every team has

Entity-Relationship Diagram (ERD)



noc_regions_csv	
ABC noc	varchar
ABC region	varchar
ABC notes	varchar

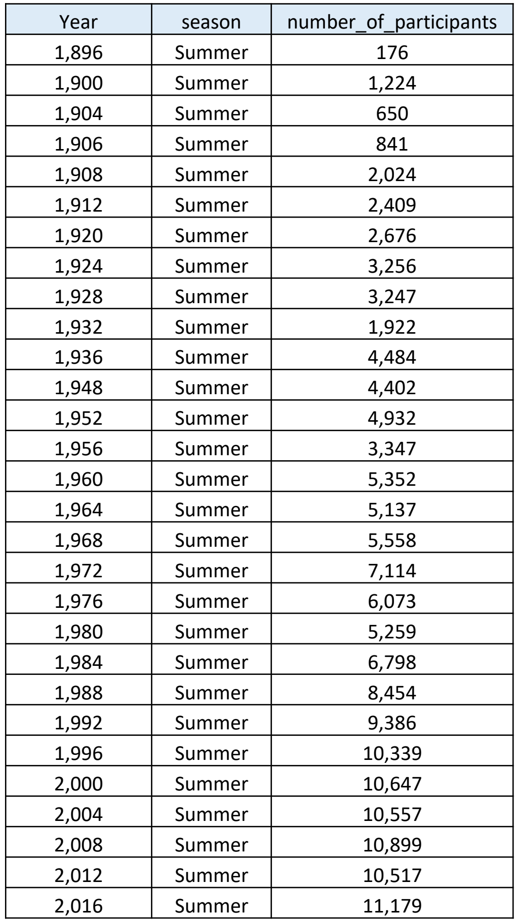
participants	
123 id	int4
ABC name	varchar
ABC sex	varchar
123 Year	int4
ABC season	varchar
ABC noc	varchar

medals	
ABC noc	varchar
123 event_id	int8
ABC medal	varchar
ABC gender	varchar
123 Year	int4
ABC season	varchar

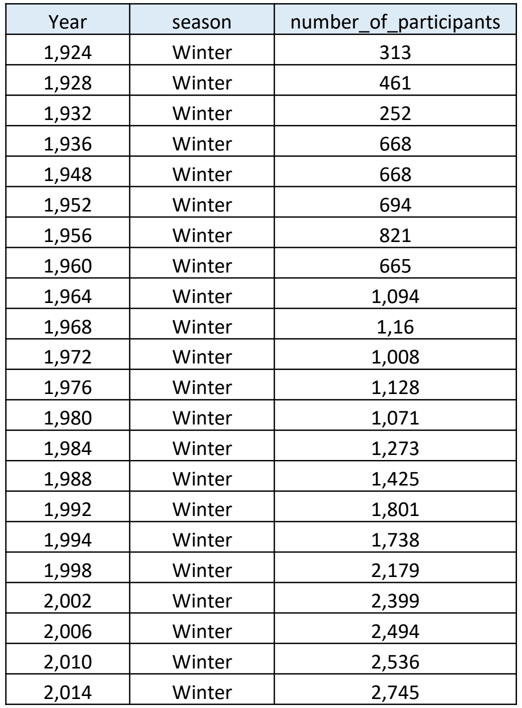
games	
ABC city	varchar
123 Year	int4
ABC season	varchar
123 participant_id	int4

sports	
123 id	int8
ABC Event	varchar
ABC sport	varchar

Number of participants in Olympic Games

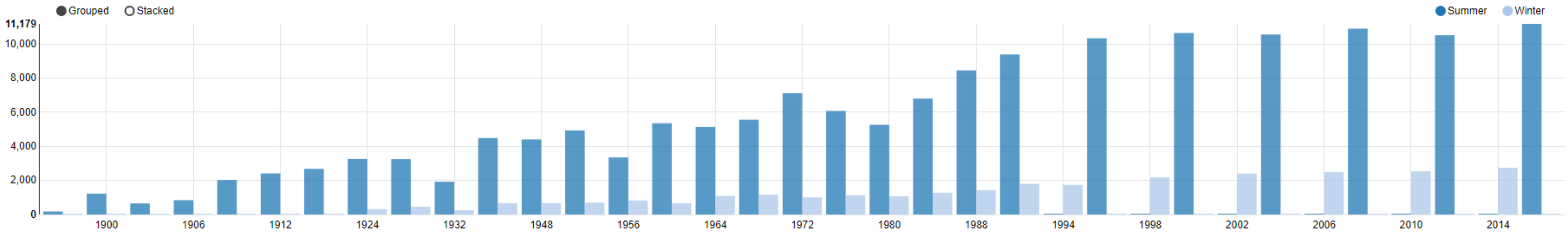


Year	season	number_of_participants
1,896	Summer	176
1,900	Summer	1,224
1,904	Summer	650
1,906	Summer	841
1,908	Summer	2,024
1,912	Summer	2,409
1,920	Summer	2,676
1,924	Summer	3,256
1,928	Summer	3,247
1,932	Summer	1,922
1,936	Summer	4,484
1,948	Summer	4,402
1,952	Summer	4,932
1,956	Summer	3,347
1,960	Summer	5,352
1,964	Summer	5,137
1,968	Summer	5,558
1,972	Summer	7,114
1,976	Summer	6,073
1,980	Summer	5,259
1,984	Summer	6,798
1,988	Summer	8,454
1,992	Summer	9,386
1,996	Summer	10,339
2,000	Summer	10,647
2,004	Summer	10,557
2,008	Summer	10,899
2,012	Summer	10,517
2,016	Summer	11,179



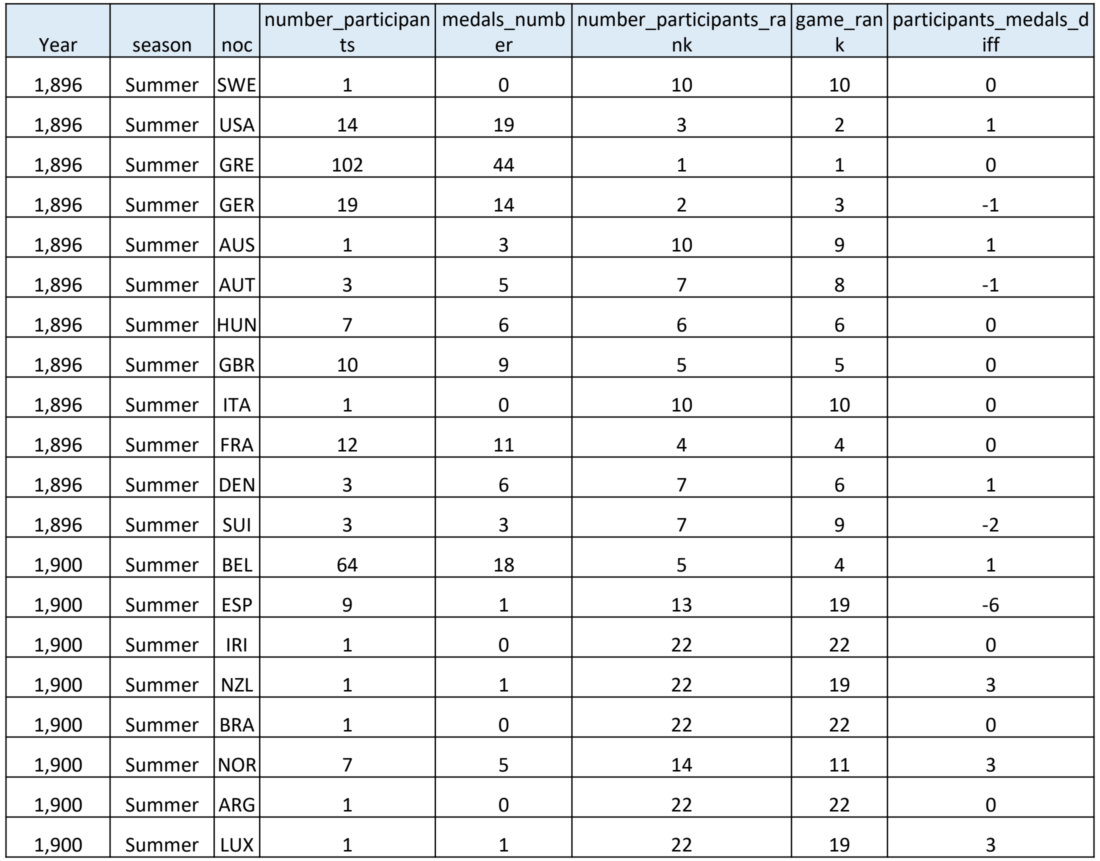
Year	season	number_of_participants
1,924	Winter	313
1,928	Winter	461
1,932	Winter	252
1,936	Winter	668
1,948	Winter	668
1,952	Winter	694
1,956	Winter	821
1,960	Winter	665
1,964	Winter	1,094
1,968	Winter	1,16
1,972	Winter	1,008
1,976	Winter	1,128
1,980	Winter	1,071
1,984	Winter	1,273
1,988	Winter	1,425
1,992	Winter	1,801
1,994	Winter	1,738
1,998	Winter	2,179
2,002	Winter	2,399
2,006	Winter	2,494
2,010	Winter	2,536
2,014	Winter	2,745

Number of participants in Olympic Games



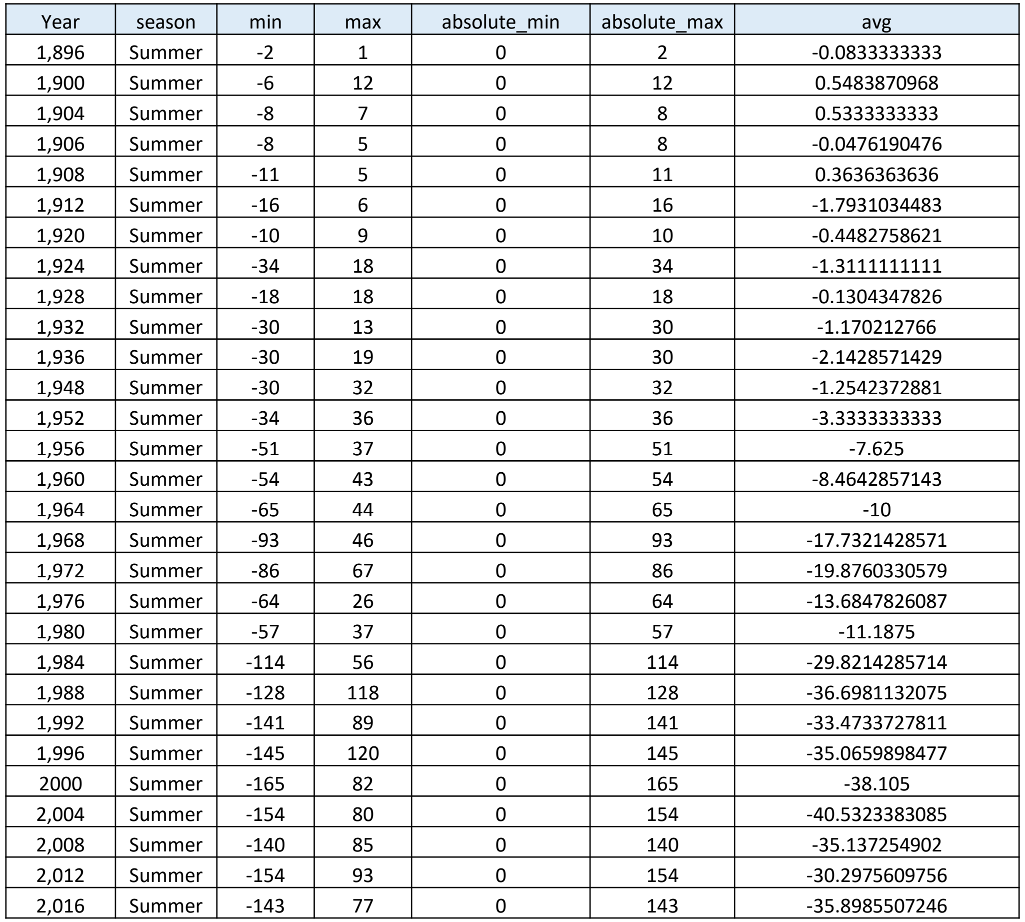
Countries and medal ranks

- I built two metrics: rank by number of participants and rank by number of medals for teams in every Olympic Games
- I checked the differences between those ranks
- *Fun fact: In 1896 Summer Olympic games there were teams that had more medals than participants. Australia had only one participant and he has got 3 medals. His name was Edwin Harold "Teddy" Flack.*



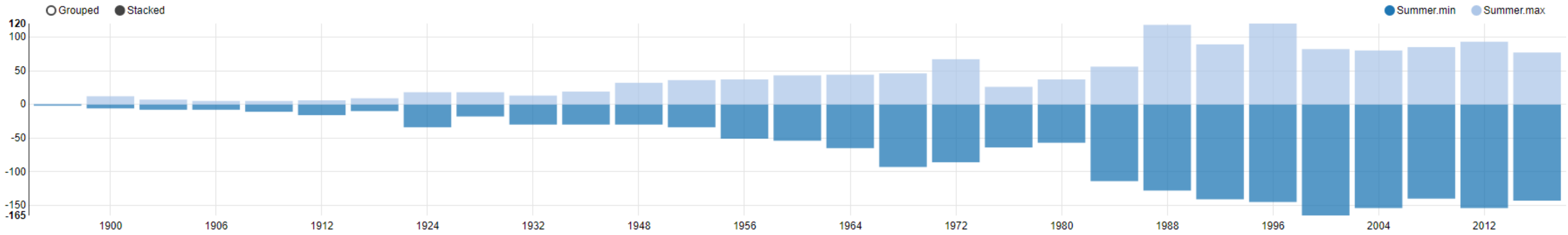
Year	season	noc	number_participants	medals_number	number_participants_rank	game_rank	participants_medals_diff
1,896	Summer	SWE	1	0	10	10	0
1,896	Summer	USA	14	19	3	2	1
1,896	Summer	GRE	102	44	1	1	0
1,896	Summer	GER	19	14	2	3	-1
1,896	Summer	AUS	1	3	10	9	1
1,896	Summer	AUT	3	5	7	8	-1
1,896	Summer	HUN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

Descriptive statistics for participants_medals_diff - Summer

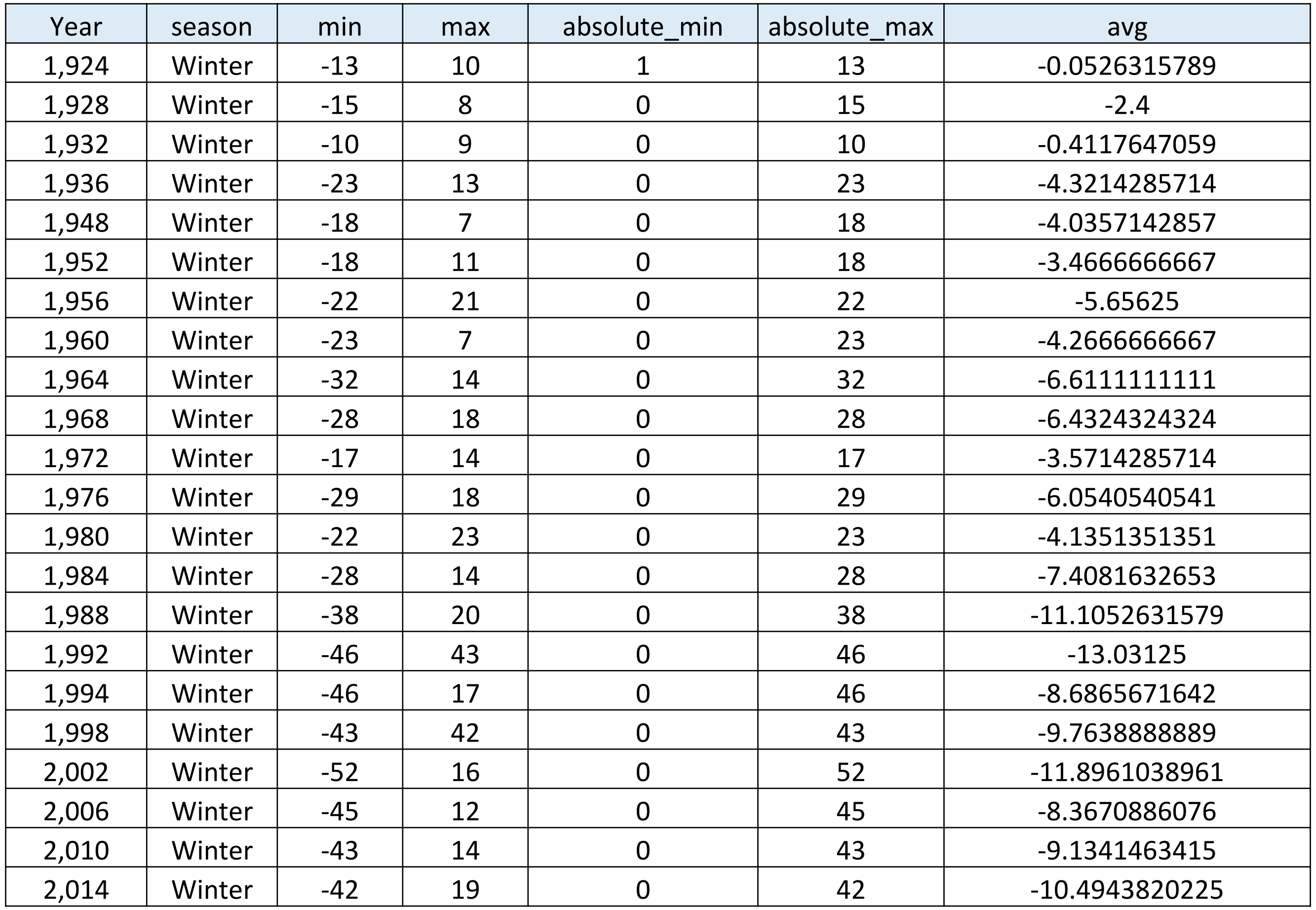


Year	season	min	max	absolute_min	absolute_max	avg
1,896	Summer	-2	1	0	2	-0.0833333333
1,900	Summer	-6	12	0	12	0.5483870968
1,904	Summer	-8	7	0	8	0.5333333333
1,906	Summer	-8	5	0	8	-0.0476190476
1,908	Summer	-11	5	0	11	0.3636363636
1,912	Summer	-16	6	0	16	-1.7931034483
1,920	Summer	-10	9	0	10	-0.4482758621
1,924	Summer	-34	18	0	34	-1.3111111111
1,928	Summer	-18	18	0	18	-0.1304347826
1,932	Summer	-30	13	0	30	-1.170212766
1,936	Summer	-30	19	0	30	-2.1428571429
1,948	Summer	-30	32	0	32	-1.2542372881
1,952	Summer	-34	36	0	36	-3.3333333333
1,956	Summer	-51	37	0	51	-7.625
1,960	Summer	-54	43	0	54	-8.4642857143
1,964	Summer	-65	44	0	65	-10
1,968	Summer	-93	46	0	93	-17.7321428571
1,972	Summer	-86	67	0	86	-19.8760330579
1,976	Summer	-64	26	0	64	-13.6847826087
1,980	Summer	-57	37	0	57	-11.1875
1,984	Summer	-114	56	0	114	-29.8214285714
1,988	Summer	-128	118	0	128	-36.6981132075
1,992	Summer	-141	89	0	141	-33.4733727811
1,996	Summer	-145	120	0	145	-35.0659898477
2000	Summer	-165	82	0	165	-38.105
2,004	Summer	-154	80	0	154	-40.5323383085
2,008	Summer	-140	85	0	140	-35.137254902
2,012	Summer	-154	93	0	154	-30.2975609756
2,016	Summer	-143	77	0	143	-35.8985507246

Descriptive statistics for participants_medals_diff - Summer

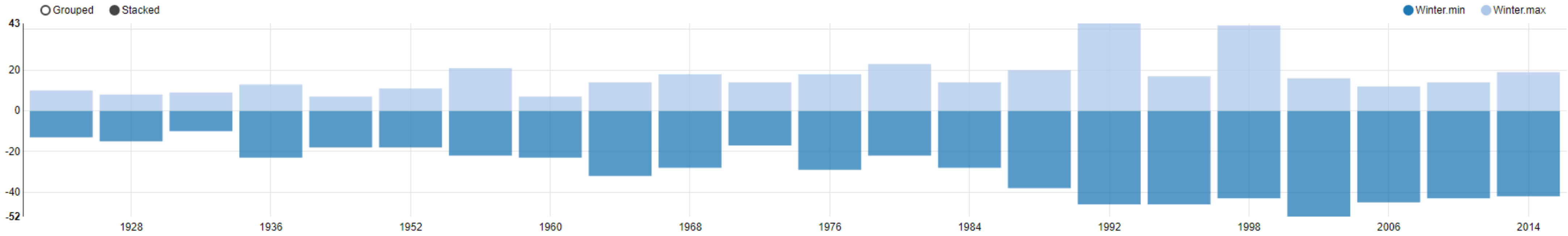


Descriptive statistics for participants_medals_diff - Winter



Year	season	min	max	absolute_min	absolute_max	avg
1,924	Winter	-13	10	1	13	-0.0526315789
1,928	Winter	-15	8	0	15	-2.4
1,932	Winter	-10	9	0	10	-0.4117647059
1,936	Winter	-23	13	0	23	-4.3214285714
1,948	Winter	-18	7	0	18	-4.0357142857
1,952	Winter	-18	11	0	18	-3.4666666667
1,956	Winter	-22	21	0	22	-5.65625
1,960	Winter	-23	7	0	23	-4.2666666667
1,964	Winter	-32	14	0	32	-6.6111111111
1,968	Winter	-28	18	0	28	-6.4324324324
1,972	Winter	-17	14	0	17	-3.5714285714
1,976	Winter	-29	18	0	29	-6.0540540541
1,980	Winter	-22	23	0	23	-4.1351351351
1,984	Winter	-28	14	0	28	-7.4081632653
1,988	Winter	-38	20	0	38	-11.1052631579
1,992	Winter	-46	43	0	46	-13.03125
1,994	Winter	-46	17	0	46	-8.6865671642
1,998	Winter	-43	42	0	43	-9.7638888889
2,002	Winter	-52	16	0	52	-11.8961038961
2,006	Winter	-45	12	0	45	-8.3670886076
2,010	Winter	-43	14	0	43	-9.1341463415
2,014	Winter	-42	19	0	42	-10.4943820225

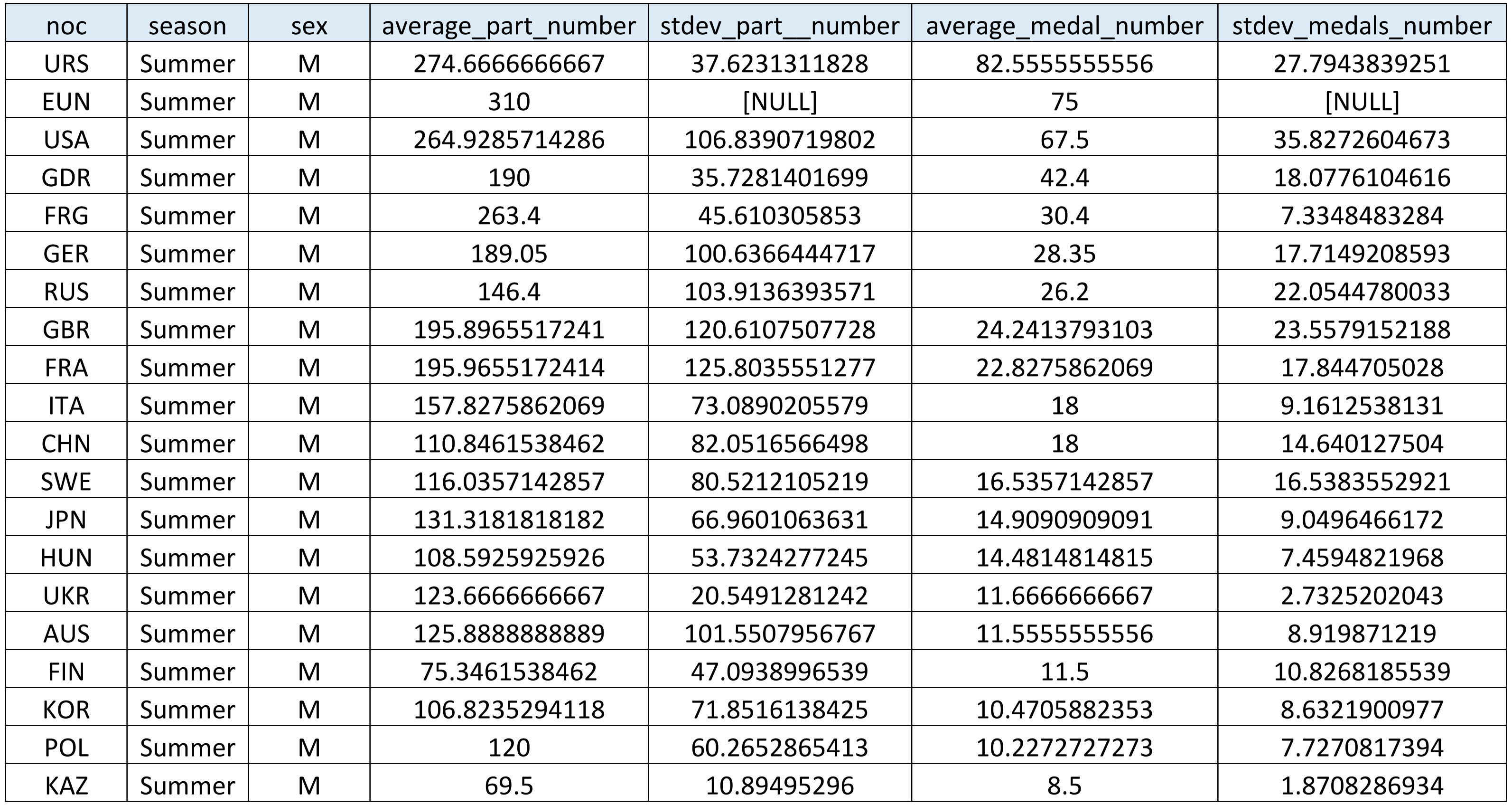
Descriptive statistics for participants_medals_diff - Winter



Descriptive statistics for participants_medals_diff - conclusion

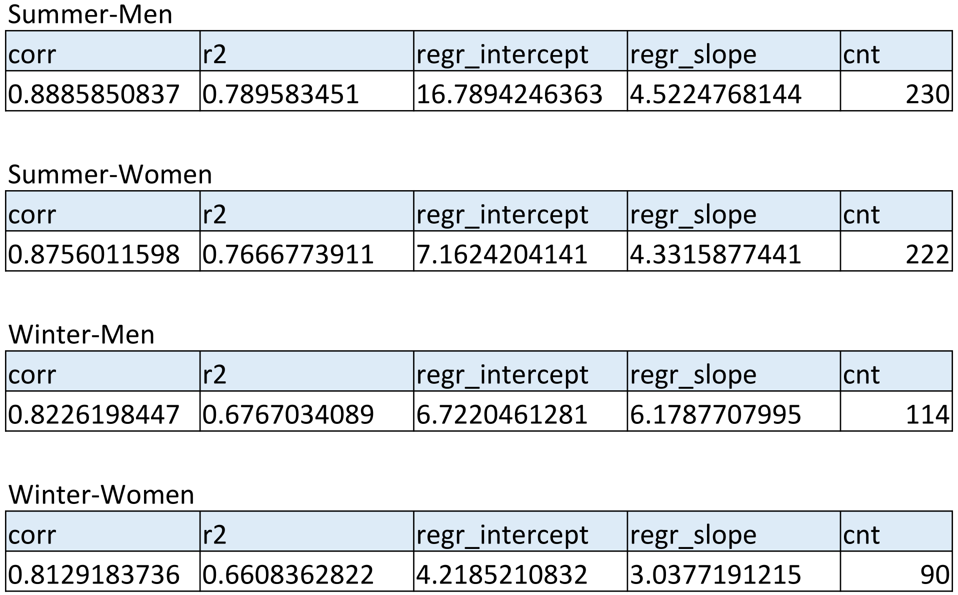
- I expected that all statistics will be close to zero. But they are not. In winter 1924 min from absolute value was one. That means that none of partcipants number and game ranks were the same.
- *Fun fact: I noticed that in 1904 the standard deviation of number of participants (and number of medals) for women in summer games was zero. That could mean that all the women teams on that games had exactly the same number of participants and got exactly the same number of medals. I checked in the data that there was only one such team, and it was USA.*

Average and standard deviation for number of medals and number of participants example for Summer-Men



noc	season	sex	average_part_number	stdev_part__number	average_medal_number	stdev_medals_number
URS	Summer	M	274.6666666667	37.6231311828	82.5555555556	27.7943839251
EUN	Summer	M	310	[NULL]	75	[NULL]
USA	Summer	M	264.9285714286	106.8390719802	67.5	35.8272604673
GDR	Summer	M	190	35.7281401699	42.4	18.0776104616
FRG	Summer	M	263.4	45.610305853	30.4	7.3348483284
GER	Summer	M	189.05	100.6366444717	28.35	17.7149208593
RUS	Summer	M	146.4	103.9136393571	26.2	22.0544780033
GBR	Summer	M	195.8965517241	120.6107507728	24.2413793103	23.5579152188
FRA	Summer	M	195.9655172414	125.8035551277	22.8275862069	17.844705028
ITA	Summer	M	157.8275862069	73.0890205579	18	9.1612538131
CHN	Summer	M	110.8461538462	82.0516566498	18	14.640127504
SWE	Summer	M	116.0357142857	80.5212105219	16.5357142857	16.5383552921
JPN	Summer	M	131.3181818182	66.9601063631	14.9090909091	9.0496466172
HUN	Summer	M	108.5925925926	53.7324277245	14.4814814815	7.4594821968
UKR	Summer	M	123.6666666667	20.5491281242	11.6666666667	2.7325202043
AUS	Summer	M	125.8888888889	101.5507956767	11.5555555556	8.919871219
FIN	Summer	M	75.3461538462	47.0938996539	11.5	10.8268185539
KOR	Summer	M	106.8235294118	71.8516138425	10.4705882353	8.6321900977
POL	Summer	M	120	60.2652865413	10.2272727273	7.7270817394
KAZ	Summer	M	69.5	10.89495296	8.5	1.8708286934

Linear regression: strong correlation between average_part_number and average_medal_number



Summer-Men

corr	r2	regr_intercept	regr_slope	cnt
0.8885850837	0.789583451	16.7894246363	4.5224768144	230

Summer-Women

corr	r2	regr_intercept	regr_slope	cnt
0.8756011598	0.7666773911	7.1624204141	4.3315877441	222

Winter-Men

corr	r2	regr_intercept	regr_slope	cnt
0.8226198447	0.6767034089	6.7220461281	6.1787707995	114

Winter-Women

corr	r2	regr_intercept	regr_slope	cnt
0.8129183736	0.6608362822	4.2185210832	3.0377191215	90

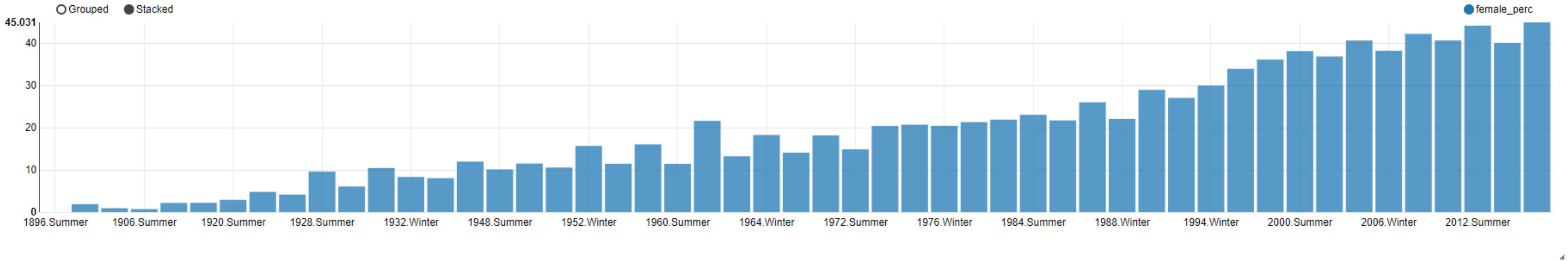
Percent of women participants growing over time

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

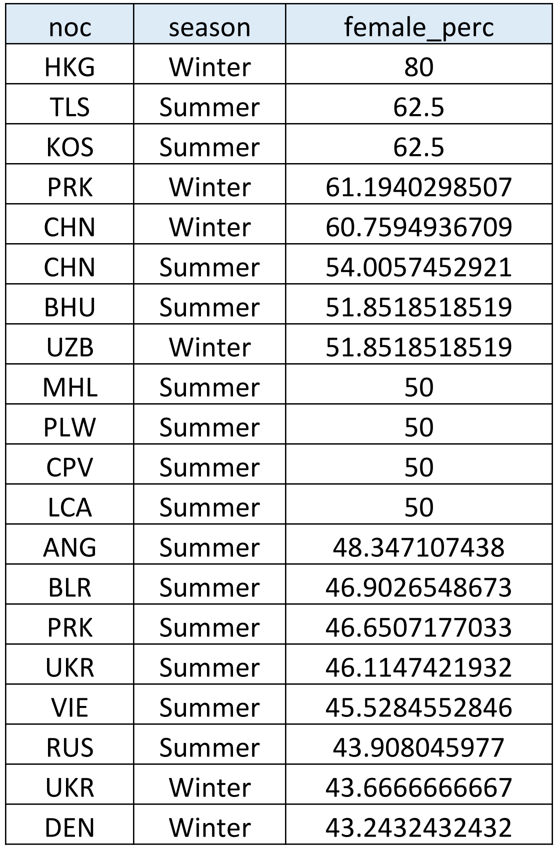
Year	season	female_perc
1,896	Summer	0
1,900	Summer	1.8790849673
1,904	Summer	0.9230769231
1,906	Summer	0.7134363853
1,908	Summer	2.1739130435
1,912	Summer	2.200083022
1,920	Summer	2.9147982063
1,924	Summer	4.7911547912
1,924	Winter	4.1533546326
...		
...		
...		
2,002	Winter	36.9320550229
2,004	Summer	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

Percent of women participants growing over time – correlation between „Year” and female_perc

corr	r2	regr_intercept	regr_slope	cnt
0.9520753909	0.90644755	1,915.8001528124	2.4633563774	51



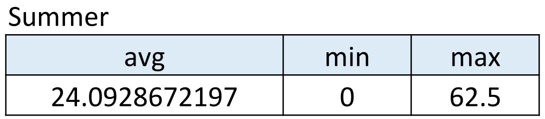
Which of the team has the highest percent of women participants



noc	season	female_perc
HKG	Winter	80
TLS	Summer	62.5
KOS	Summer	62.5
PRK	Winter	61.1940298507
CHN	Winter	60.7594936709
CHN	Summer	54.0057452921
BHU	Summer	51.8518518519
UZB	Winter	51.8518518519
MHL	Summer	50
PLW	Summer	50
CPV	Summer	50
LCA	Summer	50
ANG	Summer	48.347107438
BLR	Summer	46.9026548673
PRK	Summer	46.6507177033
UKR	Summer	46.1147421932
VIE	Summer	45.5284552846
RUS	Summer	43.908045977
UKR	Winter	43.6666666667
DEN	Winter	43.2432432432

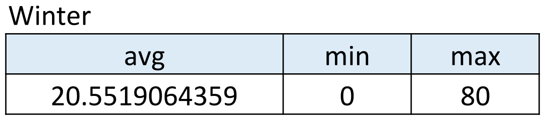
Percent of women participants in teams

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



Summer

avg	min	max
24.0928672197	0	62.5



Winter

avg	min	max
20.5519064359	0	80

Insights Discovered

- Analysis of Olympic Games dataset proves that number of participants per country is highly correlated to rank in medals.

At first I build metrics with ranks and and number of participants and I checked how they behave for olympic teams. I believed that looking at those data would give me the answers for my main hyphotesis. I would say that this approach led me to nowhere specific.

Second approach was much simpler: a checked corelation between average number of medals and average number of participants per team. They were highly correlated.

Insights Discovered

- I proved that percent of women participants is growing every year.
- 80% of participants of Hong Kong team in winter Olympic games were women
- There are some teams that never had women participants in Olympic Games

Recommendations and Actions

- In my analyses I focused on specific questions. I believe that this dataset can give us much more. I didn’t even consider information about age, height and weight of participants. I belived that there was no place for them in this specific study but there is opportunity for more analyses. Even with the data that I used you can find more questions to answer.
- This data is becoming incomplete. It should be updated with new data after every olympic games. It would be nice to know if the trends that were observed would continue. Maybe they will slow down and we would observe some kind of stagnation or maybe they would even reverse. The future will tell.