Olympics games

SQL for Data Science Capstone Project

Oslo, Norway

Itamara Campos

Milestone week 04

Review criteria

Your presentation will be a culmination of the other milestones you completed in this project-based course. You will create your presentation using any media you choose and use the Rich Text Editor feature to submit your presentation.

For presentation ideas:

- Look at DataBricks and markdown (notebooks)
- Visualizations … raw data Infographics
- Presentation Styles / Audiences
- Reference SQL output vs. visualizations

Build on Project Proposal

Build on your project proposal (from Milestone 1) that described the client or dataset you chose, the approach you were going to take, your initial hypotheses, and your initial approach. Include descriptive stats and any visualizations from your data exploration. You want to highlight key learnings from your data exploration and any aha's or changes to your plan as a results of your findings:

- Include Client/Hypotheses/Approach
- Include artifacts from previous modules
- Include results (good and bad paths); Correlations / regressions
- Graphics / Visualizations

Discuss Insights Discovered

Discuss insights discovered (results from your diving deeper / going broader analysis). This is where you put your spin on what you've discovered

- Discuss your hypotheses and any direct outcomes from whether you were right or wrong. Did you change your hypotheses? Or create new ones?
- Discuss any metrics you created and why?
- Discuss discoveries about relationships in the data / themes discovered.

Recommendations and Actions

Summarize the insights you found and make recommendations on what your client should do. What is the next steps or the action that should be taken as a result of your analysis?

Content

Hypothesis

1. Over the years women participate more in the sports, that also include gain some medals.
2. Gender equality (men and woman) must be increased in various sports.
3. New sports (modalities) in the Olympic games. Maybe an option? Need investigation.

Build on Project Proposal

- □ Target audience: everyone that likes watching the Olympic games, all genders and ages
- □ The SQL code is inside of app.model queries (link in the end of the presentation)
- □ The graphics were generated by app.mode, excel and Tableau public (final visualization it will be provided)

Discuss Insights Discovered

Based on absolute number the database show a big number of men in sport. Men are the majority, but includes an analyze over time, we can see the woman curve is more incline, shows an increased number of woman in the Olympic games That prove hypothesis number 1. Split the graphic between men and woman over time in various sport we can see a similarity between the graphic. That prove hypothesis number 2. Making an analyze of sport over time, plotting dots we can new dots after 1990, prove hypothesis number 3.

Recommendations and Actions

I will recommend:

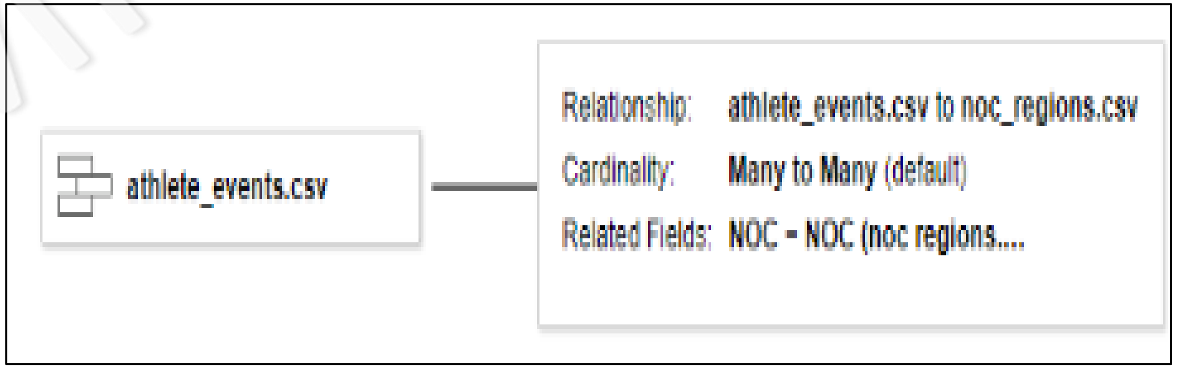
- who might be interested in this data analyses
- how this data might be useful for market proposes

Content

Challenges

1. **Upload the data** to suitable “*interface*” that could allow me to performance queries:
   - ➢ **Kaggle** by Google: easy to upload the data. But require some knowledge in Python to performance the queries. But aif using as an overview of the table that is good enough.
   - ➢ **Databricks community edition** has a lot videos show how to use. I have problem to create a Cluster, it was taking hours. Which implicit a lot time just waiting for the internal system to work. Couldn’t afford to waste time.
   - ➢ **App.mode**: It has two ways to upload the files: I) using notebook and Python language, I tried didn't work for me. II) o ask for support to upload into the database. It works. Not necessary use python language. Can easy performance the SQL queries
   - ➢ **Tableau Desktop** (local PC): tableau as a connection with SQL database. Which makes simple to performance and do graphics at the same time. It is not for free. But I could public the final visualization using Tableau Public

2. **Looking into the database.**
   Tableau Desktop can build the ERD diagram (image in the left)



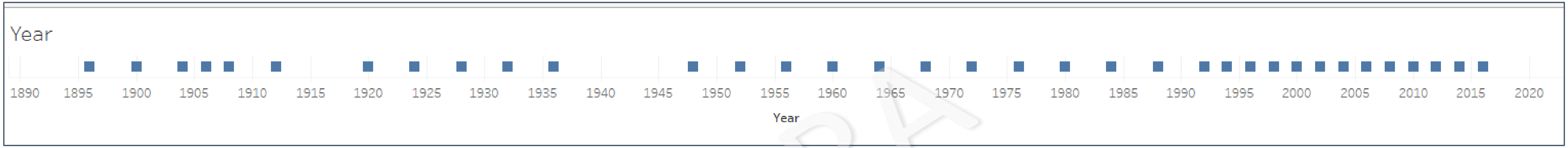
Relationship: athlete_events.csv to noc_regions.csv
Cardinality: Many to Many (default)
Related Fields: NOC = NOC (noc regions....

3. **Analise the database**
   Events columns should be called as subcategory of the sport.
   That could be misleading to false conclusion
   id columns “looks” repeat but is not. The same *id* participated of different events in the same year.
   That could be misleading to false conclusion

Woman in the Olympics Games

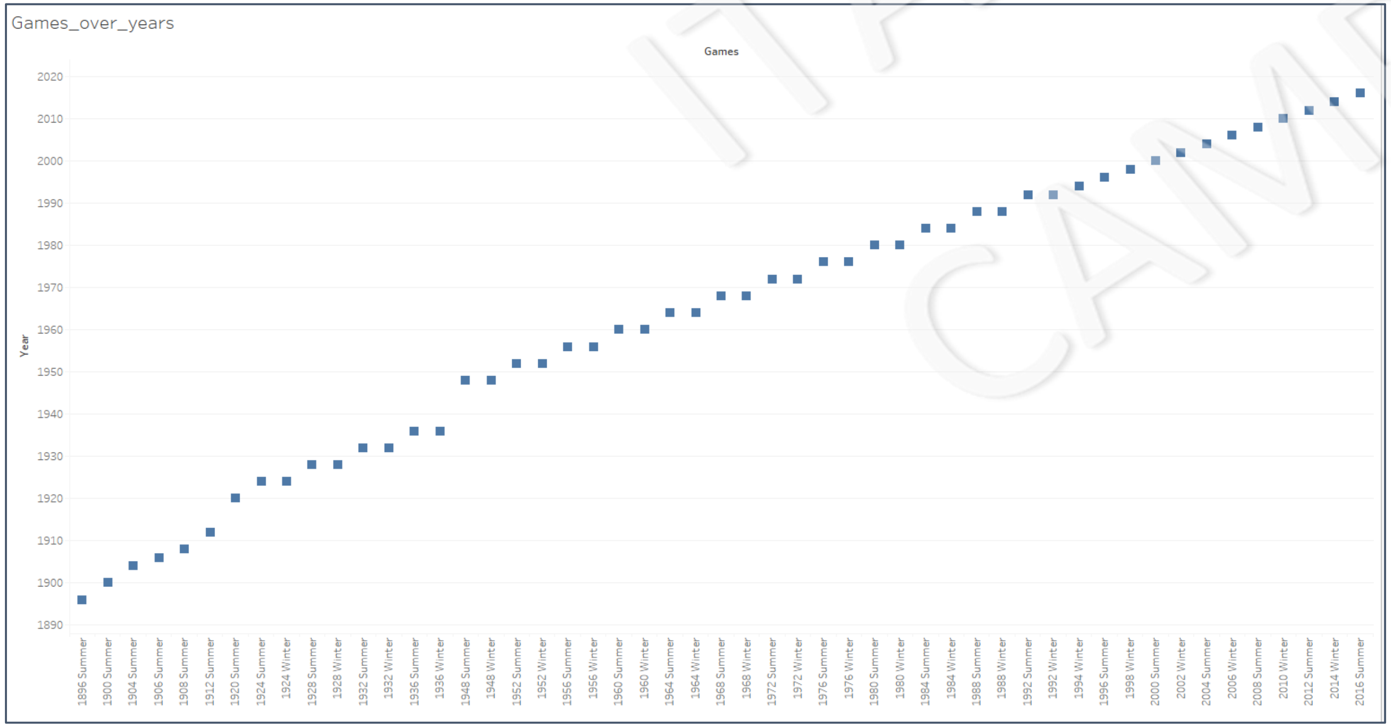
An Analise over time

Olympic Games overview



1. The distribution of the Olympic games wasn’t regular.
   - The gap in ~1915 might related to the I World War
   - The gar in ~1940 might related to the II World War
     - The data itself can't answers this hypotheses.

2. The distribution of the Olympic games has regular. After 1990.
   - After the II world war, the games were every 4 years.
   - After 1990 the events became after 2 years.
     - The reason can’t be answers using only this dataset.



3. The distribution of the Olympic games over season.

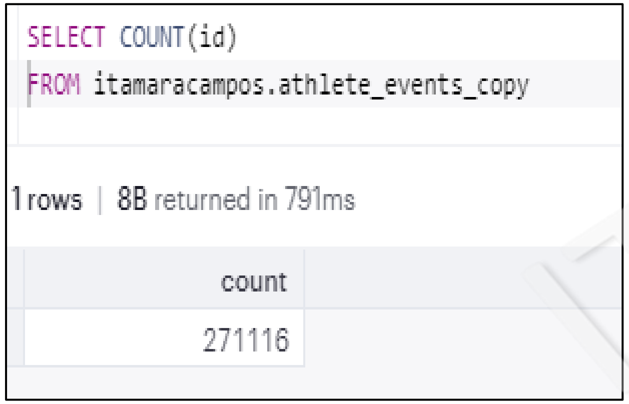
Before 1990, in the year that have the Olympic games, used to have 2 events (summer and winter) in the same year. After 1990, the games became alternated including the season.

Which show more linear in the graphic after 1990.

Statistics

Provide a summary of the different descriptive statistics you looked at and WHY.

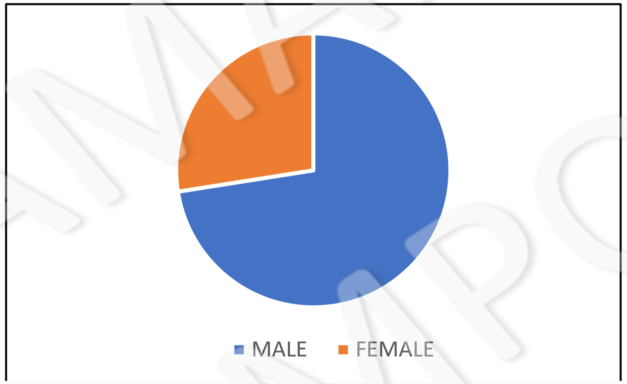
I look how many ***id*** has been in the table over years. To get an overview of the dimension of the table. (fist picture – Left)
After I look how may has ***MALE*** and ***FEMALE*** in the table. (Middle graphic – Pizza)
After a look how many has ***medals.*** I believe medal it will be a way the describe success of the statistics for all genders (Pizza right)



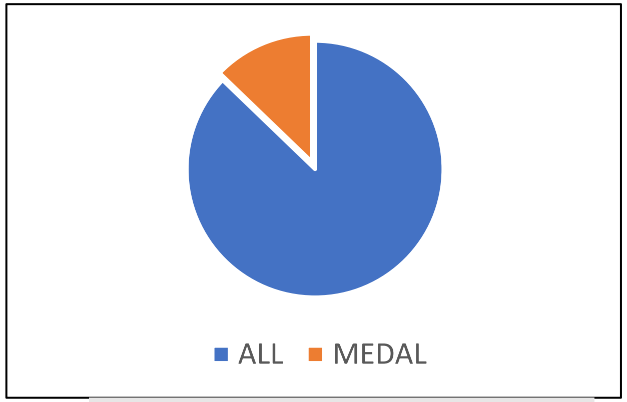
```
SELECT COUNT(id)
FROM itamaracampos.athlete_events_copy
```

```
1 rows | 8B returned in 791ms

   count
  271116
```



PORCENTAGERS	
72.51287272 %	27.48712728 %
MAN	FEMALE

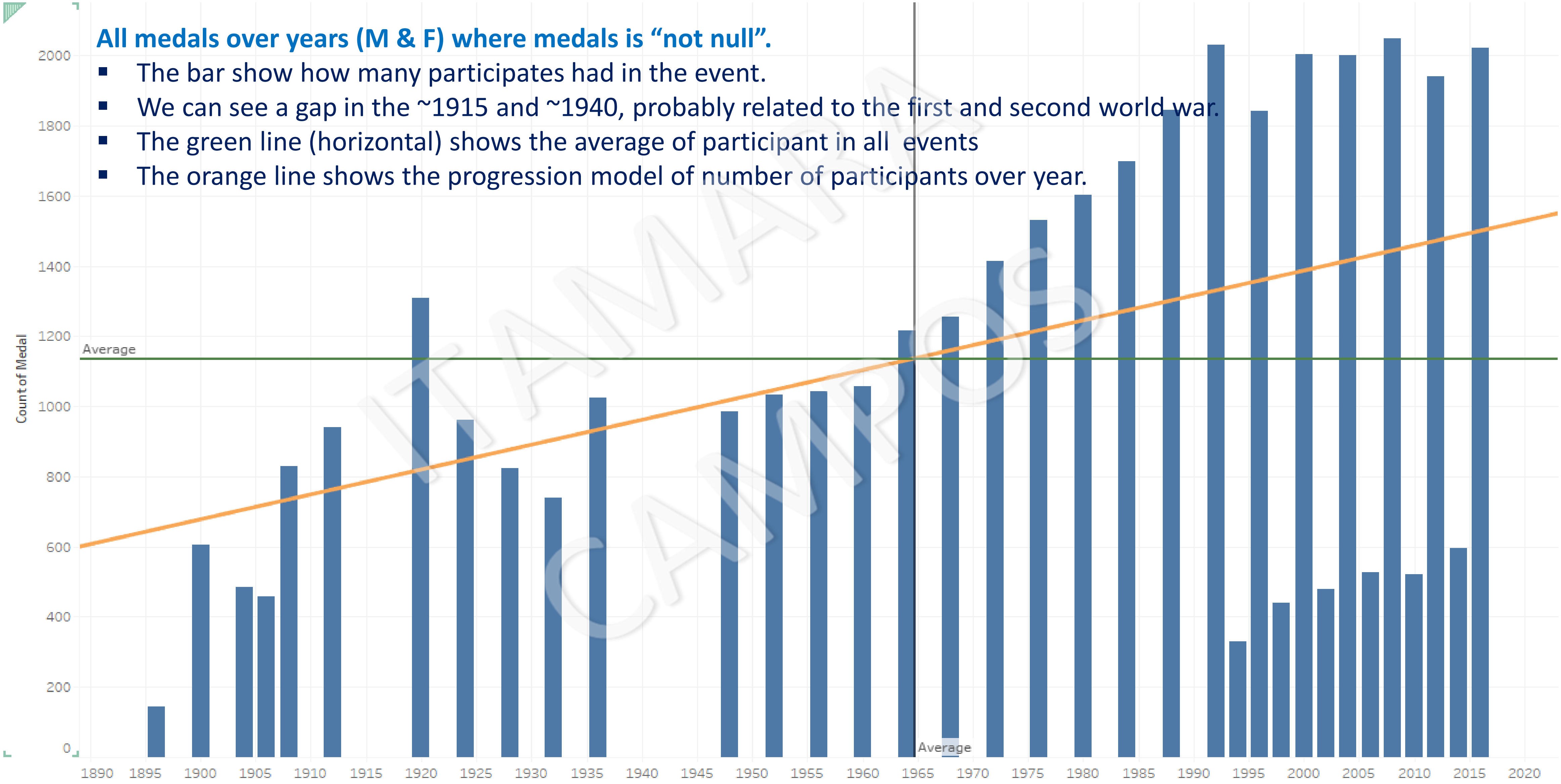


14.67379% gain medals

1. I believe over the years more people participated of the Olympiads game (see next slide analyze)

2. The gender equality (men and woman) must be increased in various sports, but I can’t tell which one. (but I believe there are more women in the events overall)

Histogram of number medal over 120 years of the Olympic Games.



All medals over years (M & F) where medals is “not null”.

- The bar show how many participates had in the event.
- We can see a gap in the ~1915 and ~1940, probably related to the first and second world war.
- The green line (horizontal) shows the average of participant in all events
- The orange line shows the progression model of number of participants over year.

This Graphic shows the number of participates in the Olympic games has been increase over years that has gain medals
But we are interested in woman in the sports over's year.

Linear regression models for male and female (split in season) over years.

Linear regression is a way of demonstrating a relationship between a dependent variable (y) and one or more explanatory variables (x). For example, on a scatterplot, linear regression finds the best fitting straight line through the data points. It is used to identify causal relationships, forecasting trends and forecasting an effect. The line of best fit comprises analyzing the correlation, and direction of the data; estimating the model; and evaluating the validity of the model.

The regression line is calculated by finding the minimized sum of squared errors of prediction. In order to calculate a straight line, you need a linear equation i.e.:

y = Mx + b

Where M= the slope of the line, b= the y-intercept and x and y are the variables. Therefore, to calculate linear regression in Tableau you first need to calculate the slope and y-intercept.

In tableau

The P-value and R-squared are vital when it comes to assessing whether the trend line model is useful or not and which model is best suited to your data.

P-value

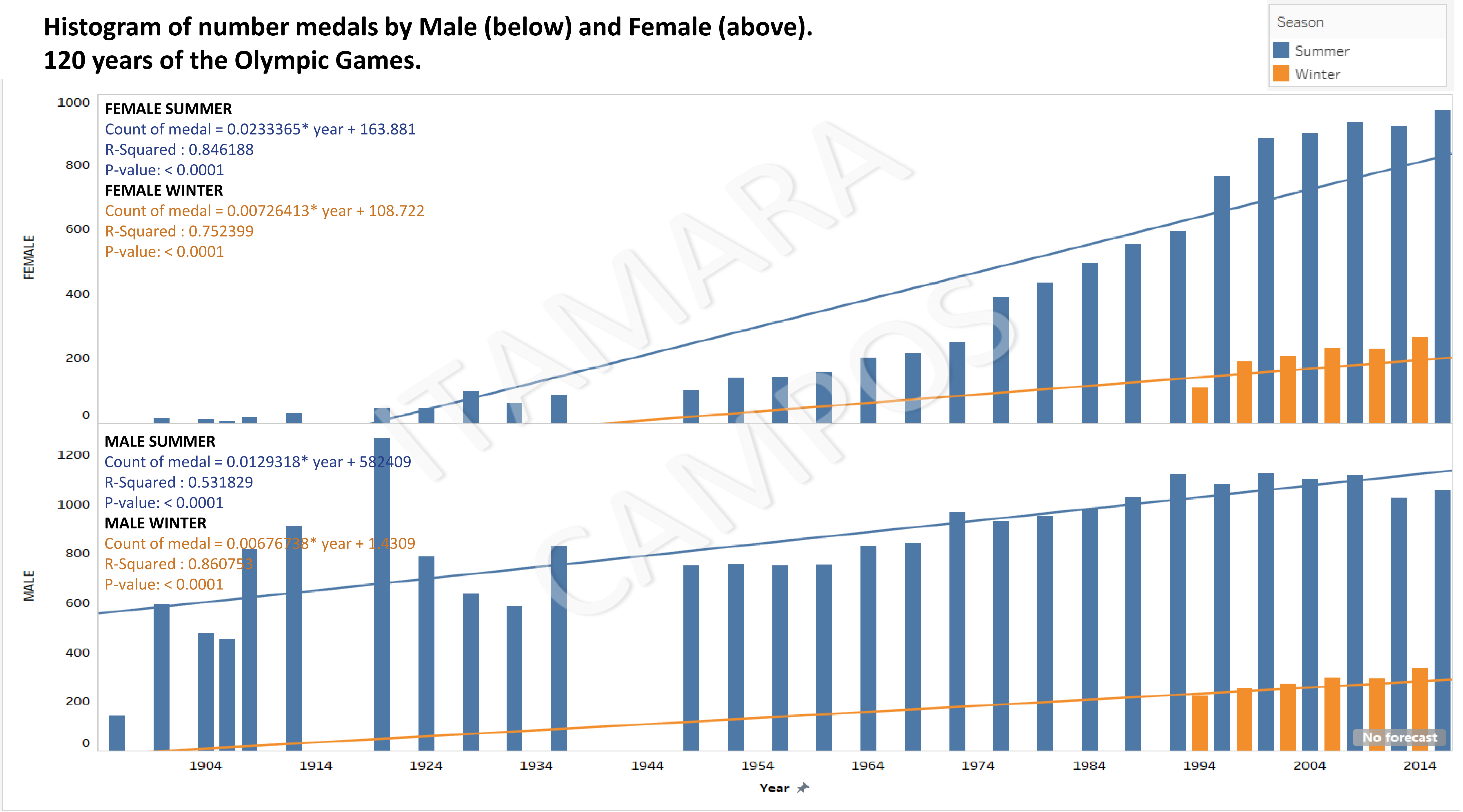
The p-value is a measure of significance for the trend line. A p-value of 0.05 or less is often considered significant; the smaller the p-value the more significant the model is. A large p-value can indicate that the apparent trend in the data is due to chance, not the factors in the model.

R-squared

The R-squared is also an important measure when assessing if the model is suitable and tells us whether the model effectively fits our data. The R-squared is measured on a scale from 0-1; the closer to 1 the more effective the model.

I used this model to analyzes two graphic (bar plot over time) with a regression model line overlaid.

Histogram of number medals by Male (below) and Female (above).
120 years of the Olympic Games.



FEMALE SUMMER

Count of medal = 0.0233365* year + 163.881

R-Squared : 0.846188

P-value: < 0.0001

FEMALE WINTER

Count of medal = 0.00726413* year + 108.722

R-Squared : 0.752399

P-value: < 0.0001

MALE SUMMER

Count of medal = 0.0129318* year + 582409

R-Squared : 0.531829

P-value: < 0.0001

MALE WINTER

Count of medal = 0.00676738* year + 1.4309

R-Squared : 0.860753

P-value: < 0.0001

Histogram of number medals by Male (below) and Female (above).
Since 1990, Olympic Games.

The graphic in the previous slide show.

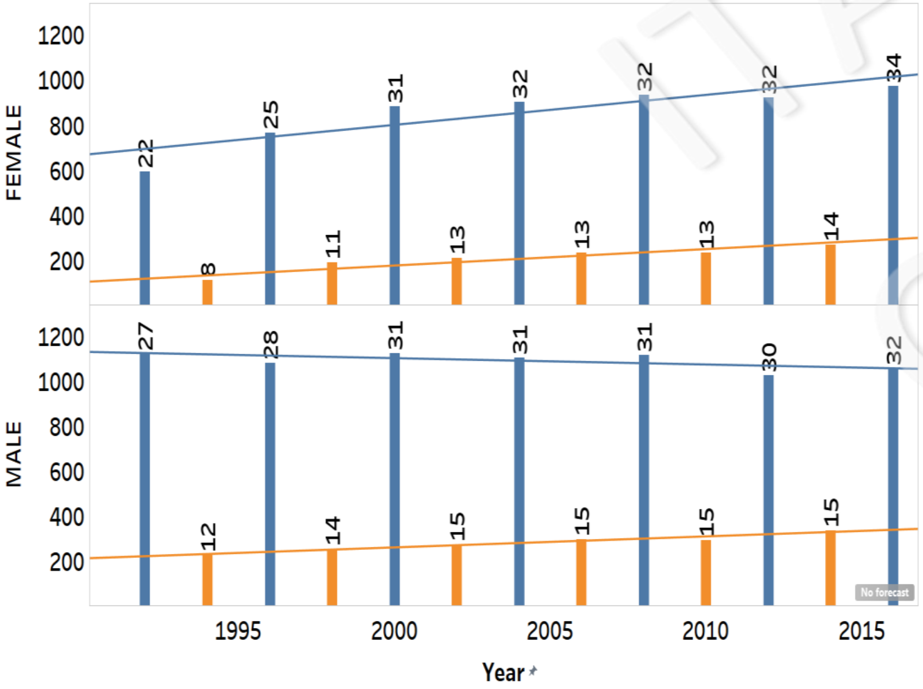
❑ The alternation of winter and summer sport after 1990

❑ Splitting the graphic by MALE and FEMALE we clearly see the woman become more active in the Olympic Games, since the graphic includes only medals non-null.

➢ Proves the hypotheses number 1

"Over the years women participate more in the sports, that also include gain some medals" ✓

Analise of one part of the graphic, see below, zoom only after 1990, because is easy to analyze.



The graphic is the number of medals (non-null) over years, split by season, classification by gender:

The number above the graphic bar is the number of sport competed in this season. Base on the number itself and in the progression line on the model we easy can say the number of woman in various sports is became almost equal to the men.

➢ Proves the hypotheses number 2

"Gender equality (men and woman) must be increased in various sports." ✓

Without necessary looking the graphic is for certain the woman participate more of the Olympic games the man. Even thought the man curve show a crescent line. Woman in sport became more popular over the years, according to our model.

FEMALE SUMMER

Count of medal = 0.0233365* year + 163.881

R-Squared : 0.846188

P-value: < 0.0001

FEMALE WINTER

Count of medal = 0.00726413* year + 108.722

R-Squared : 0.752399

P-value: < 0.0001

MALE SUMMER

Count of medal = 0.0129318* year + 582409

R-Squared : 0.531829

P-value: < 0.0001

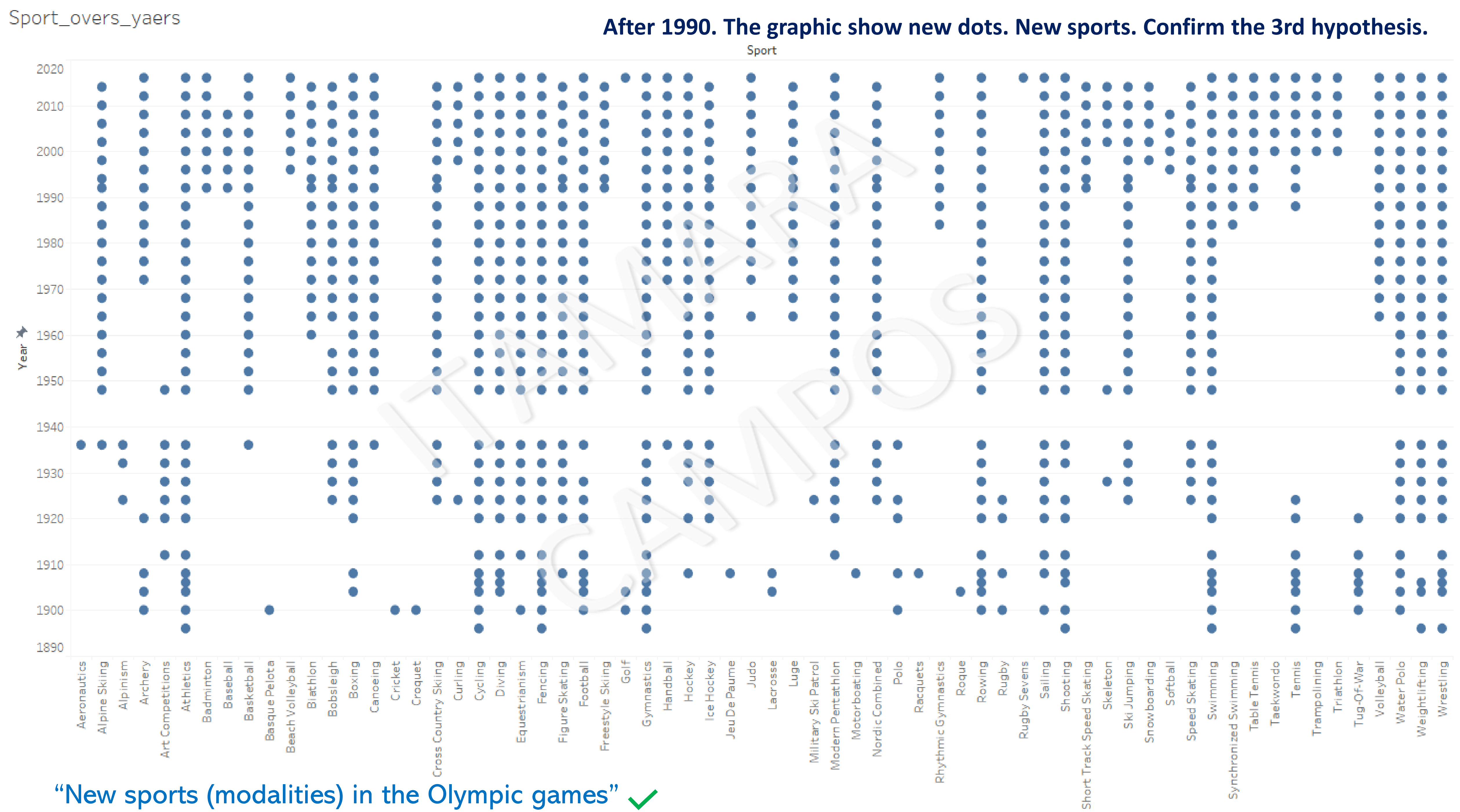
MALE WINTER

Count of medal = 0.00676738* year + 1.4309

R-Squared : 0.860753

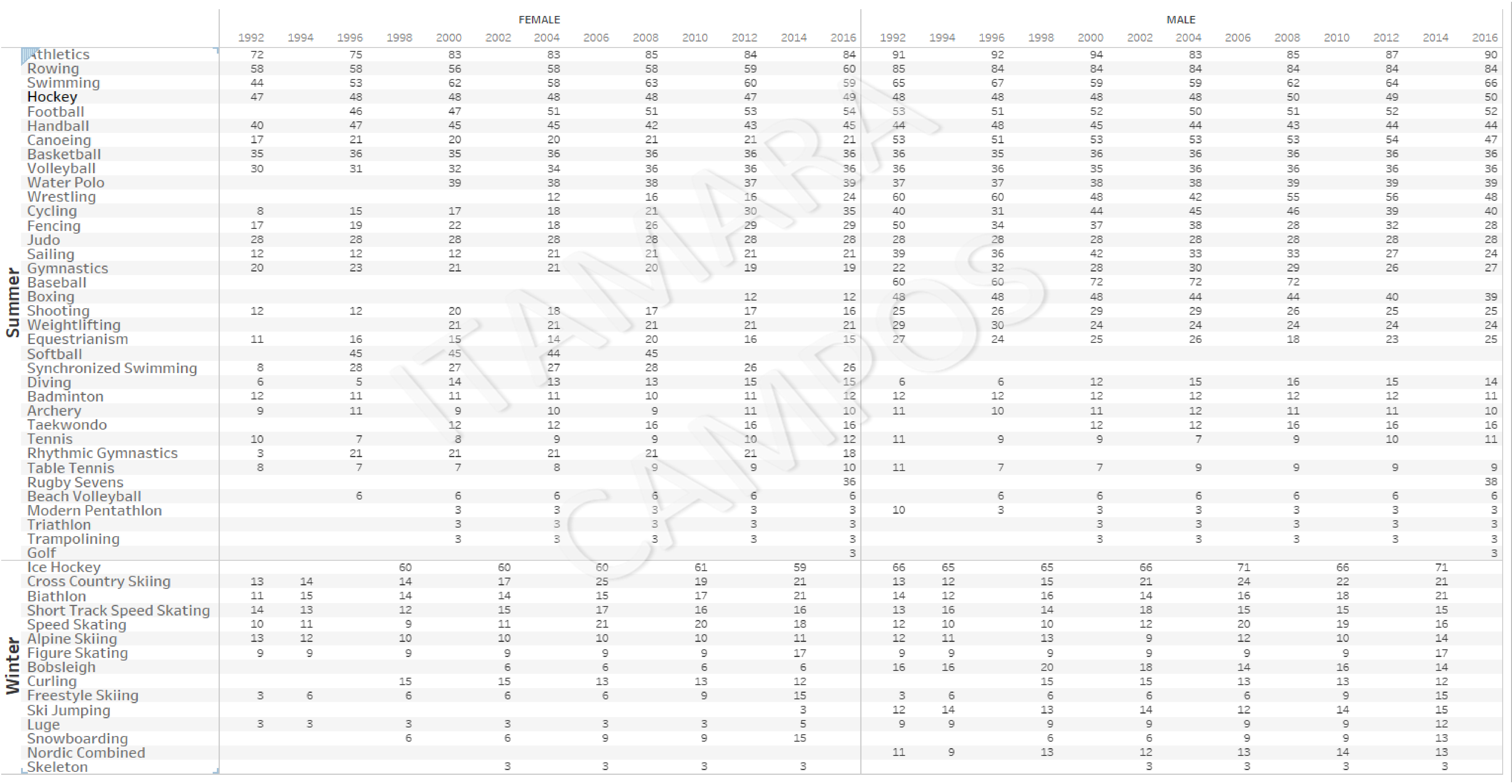
P-value: < 0.0001

After 1990. The graphic show new dots. New sports. Confirm the 3rd hypothesis.



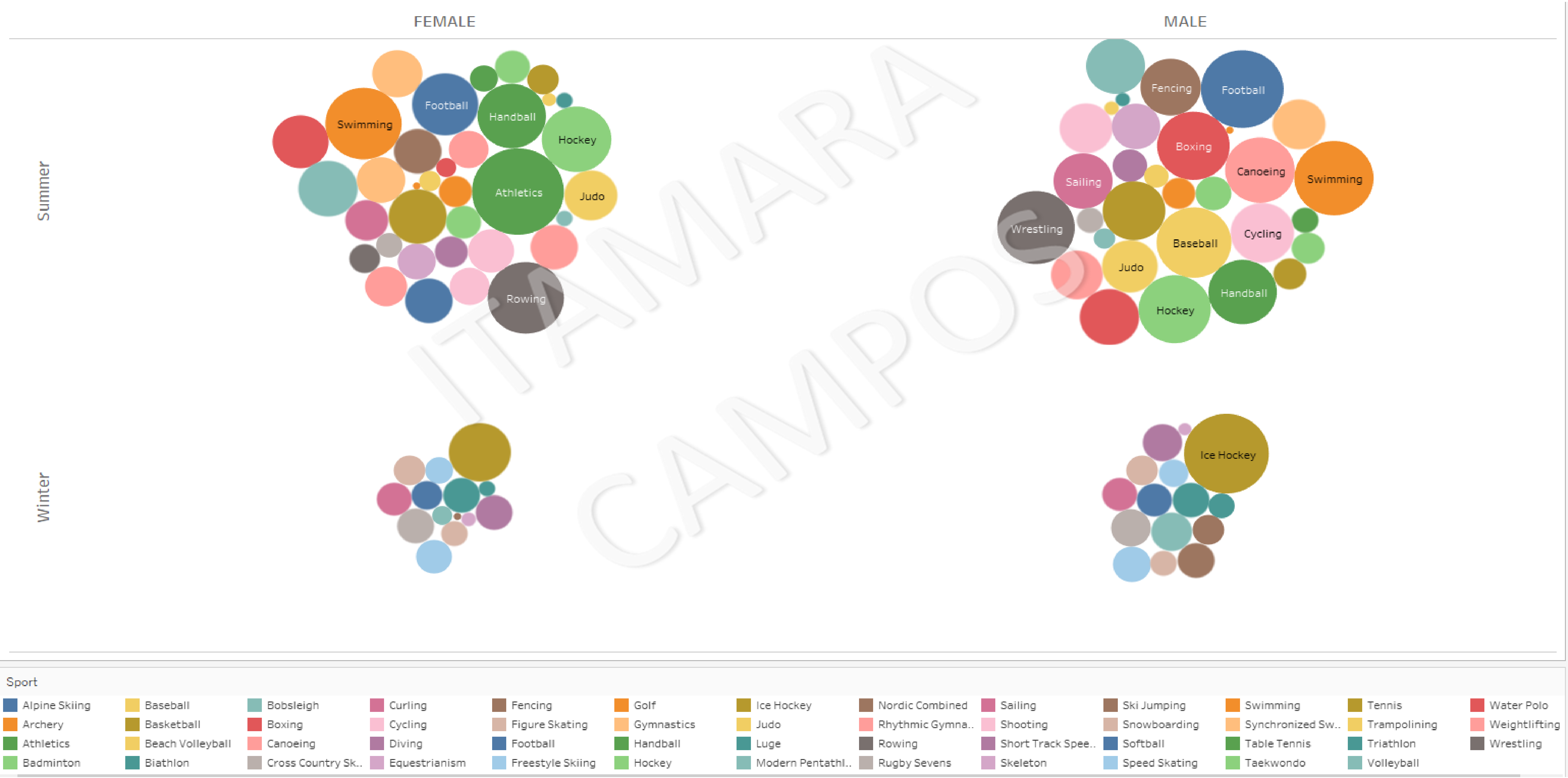
“New sports (modalities) in the Olympic games” ✓

After 1990. Table split by sport, season and sex. The number show the number of participants of the sport in the respective year.



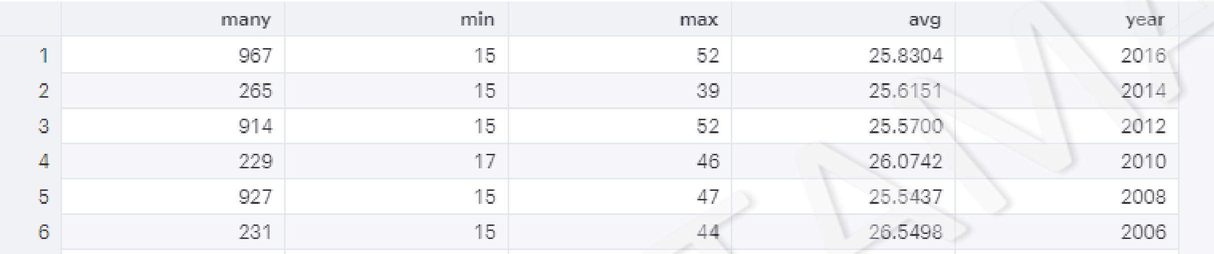
		FEMALE													MALE												
		1992	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014	2016	1992	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014	2016
Summer	Athletics	72		75		83		83		85		84		84	91		92		94		83		85		87		90
	Rowing	58		58		56		58		58		59		60	85		84		84		84		84		84		84
	Swimming	44		53		62		58		63		60		59	65		67		59		59		62		64		66
	Hockey	47		48		48		48		48		47		49	48		48		48		48		50		49		50
	Football			46		47		51		51		53		54	53		51		52		50		51		52		52
	Handball	40		47		45		45		42		43		45	44		48		45		44		43		44		44
	Canoeing	17		21		20		20		21		21		21	53		51		53		53		53		54		47
	Basketball	35		36		35		36		36		36		36	36		35		36		36		36		36		36
	Volleyball	30		31		32		34		36		36		36	36		36		35		36		36		36		36
	Water Polo					39		38		38		37		39	37		37		38		38		39		39		39
	Wrestling							12		16		16		24	60		60		48		42		55		56		48
	Cycling	8		15		17		18		21		30		35	40		31		44		45		46		39		40
	Fencing	17		19		22		18		26		29		29	50		34		37		38		28		32		28
	Judo	28		28		28		28		28		28		28	28		28		28		28		28		28		28
	Sailing	12		12		12		21		21		21		21	39		36		42		33		33		27		24
	Gymnastics	20		23		21		21		20		19		19	22		32		28		30		29		26		27
	Baseball														60		60		72		72		72				
	Boxing											12		12	48		48		48		44		44		40		39
	Shooting	12		12		20		18		17		17		16	25		26		29		29		26		25		25
	Weightlifting					21		21		21		21		21	29		30		24		24		24		24		24
	Equestrianism	11		16		15		14		20		16		15	27		24		25		26		18		23		25
	Softball			45		45		44		45																	
	Synchronized Swimming	8		28		27		27		28		26		26													
	Diving	6		5		14		13		13		15		15	6		6		12		15		16		15		14
	Badminton	12		11		11		11		10		11		12	12		12		12		12		12		12		11
	Archery	9		11		9		10		9		11		10	11		10		11		12		11		11		10
	Taekwondo					12		12		16		16		16					12		12		16		16		16
	Tennis	10		7		8		9		9		10		12	11		9		9		7		9		10		11
	Rhythmic Gymnastics	3		21		21		21		21		21		18													
	Table Tennis	8		7		7		8		9		9		10	11		7		7		9		9		9		9
	Rugby Sevens													36													38
	Beach Volleyball			6		6		6		6		6		6			6		6		6		6		6		6
	Modern Pentathlon					3		3		3		3		3	10		3		3		3		3		3		3
	Triathlon					3		3		3		3		3					3		3		3		3		3
	Trampolining					3		3		3		3		3					3		3		3		3		3
	Golf													3													3
Winter	Ice Hockey				60		60		60		61		59		66	65		65		66		71		66		71	
	Cross Country Skiing	13	14		14		17		25		19		21		13	12		15		21		24		22		21	
	Biathlon	11	15		14		14		15		17		21		14	12		16		14		16		18		21	
	Short Track Speed Skating	14	13		12		15		17		16		16		13	16		14		18		15		15		15	
	Speed Skating	10	11		9		11		21		20		18		12	10		10		12		20		19		16	
	Alpine Skiing	13	12		10		10		10		10		11		12	11		13		9		12		10		14	
	Figure Skating	9	9		9		9		9		9		17		9	9		9		9		9		9		17	
	Bobsleigh						6		6		6		6		16	16		20		18		14		16		14	
	Curling				15		15		13		13		12					15		15		13		13		12	
	Freestyle Skiing	3	6		6		6		6		9		15		3	6		6		6		6		9		15	
	Ski Jumping												3		12	14		13		14		12		14		15	
	Luge	3	3		3		3		3		3		5		9	9		9		9		9		9		12	
	Snowboarding				6		6		9		9		15					6		6		9		9		13	
	Nordic Combined														11	9		13		12		13		14		13	
	Skeleton						3		3		3		3							3		3		3		3	

After 1990. Package babble split by sport, season and sex. (same information as the table in the previous slide) but it is easy to see which sport is most common amount the athletes

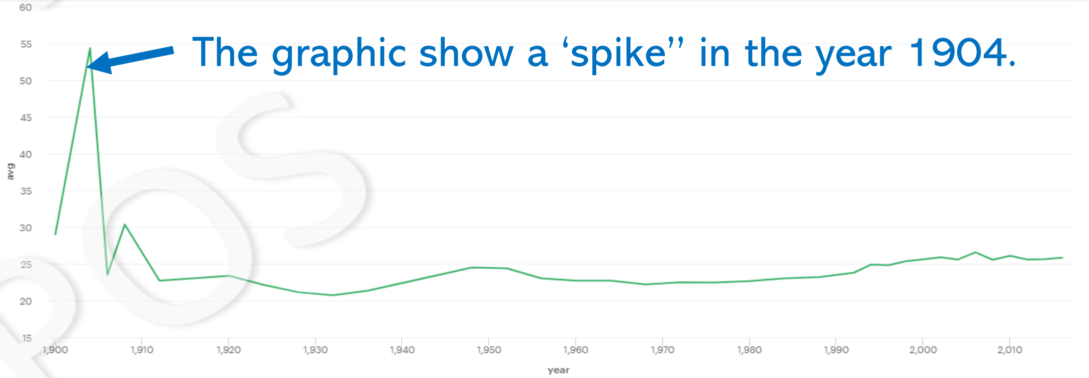


Average of woman age in the Olympic games.

```
SELECT
   COUNT(id) as MANY, Min(age), Max(age), avg(age), year
FROM
  itamaracampos.noc_regions inner JOIN itamaracampos.athlete_events_copy on itamaracampos.noc_regions.noc = itamaracampos.athlete_events_copy.noc
WHERE medal is not NULL and sex = 'F’
GROUP by year
ORDER by year DESC
```

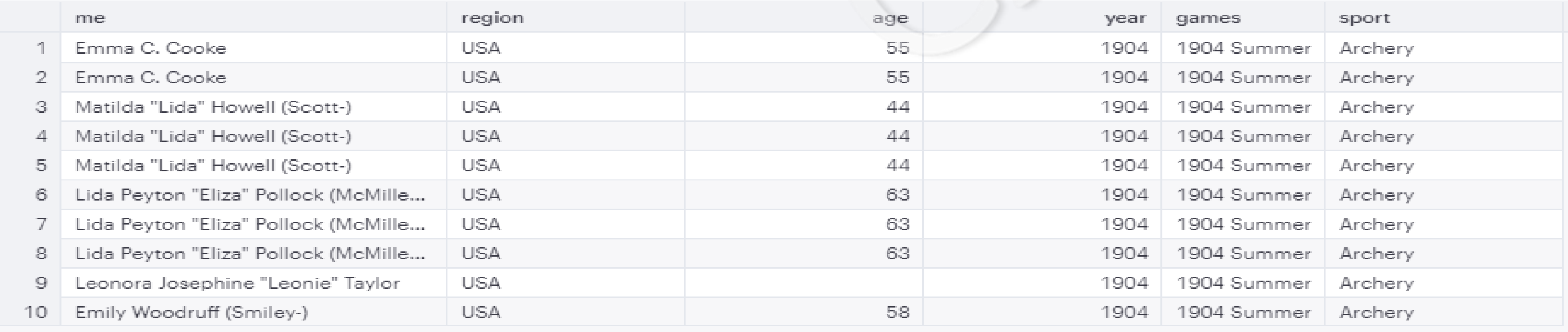


	many	min	max	avg	year
1	967	15	52	25.8304	2016
2	265	15	39	25.6151	2014
3	914	15	52	25.5700	2012
4	229	17	46	26.0742	2010
5	927	15	47	25.5437	2008
6	231	15	44	26.5498	2006



The graphic show a ‘spike’’ in the year 1904.

```
SELECT me, region, age, year, games, sportF
ROM  --itamaracampos.athlete_events_copy     itamaracampos.noc_regions inner JOIN itamaracampos.athlete_events_copy on itamaracampos.noc_regions.noc =
itamaracampos.athlete_events_copy.noc
WHERE medal is not NULL and sex = 'F' and year ='1904'
```

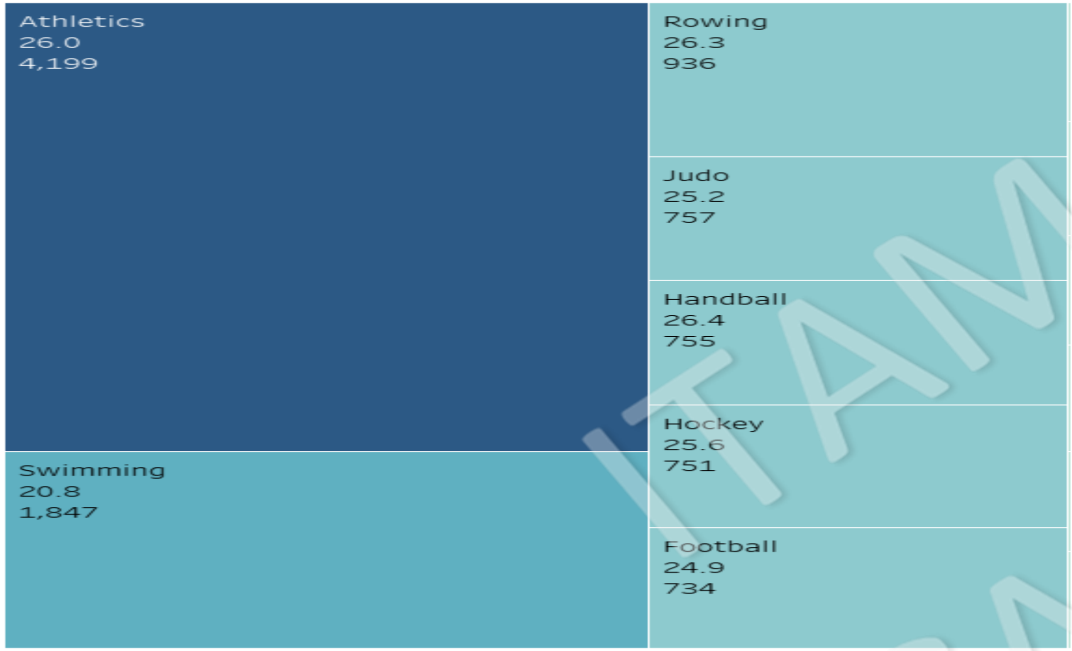


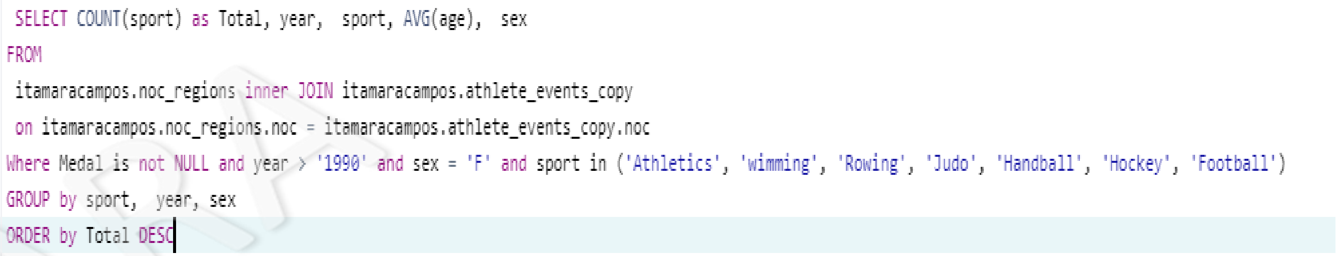
	me	region	age	year	games	sport
1	Emma C. Cooke	USA	55	1904	1904 Summer	Archery
2	Emma C. Cooke	USA	55	1904	1904 Summer	Archery
3	Matilda "Lida" Howell (Scott-)	USA	44	1904	1904 Summer	Archery
4	Matilda "Lida" Howell (Scott-)	USA	44	1904	1904 Summer	Archery
5	Matilda "Lida" Howell (Scott-)	USA	44	1904	1904 Summer	Archery
6	Lida Peyton "Eliza" Pollock (McMille...	USA	63	1904	1904 Summer	Archery
7	Lida Peyton "Eliza" Pollock (McMille...	USA	63	1904	1904 Summer	Archery
8	Lida Peyton "Eliza" Pollock (McMille...	USA	63	1904	1904 Summer	Archery
9	Leonora Josephine "Leonie" Taylor	USA		1904	1904 Summer	Archery
10	Emily Woodruff (Smiley-)	USA	58	1904	1904 Summer	Archery

There are nothing that show that is not legitime data. So still valid.
Since is only one dote. Don’t compromise the rest of the Analise.

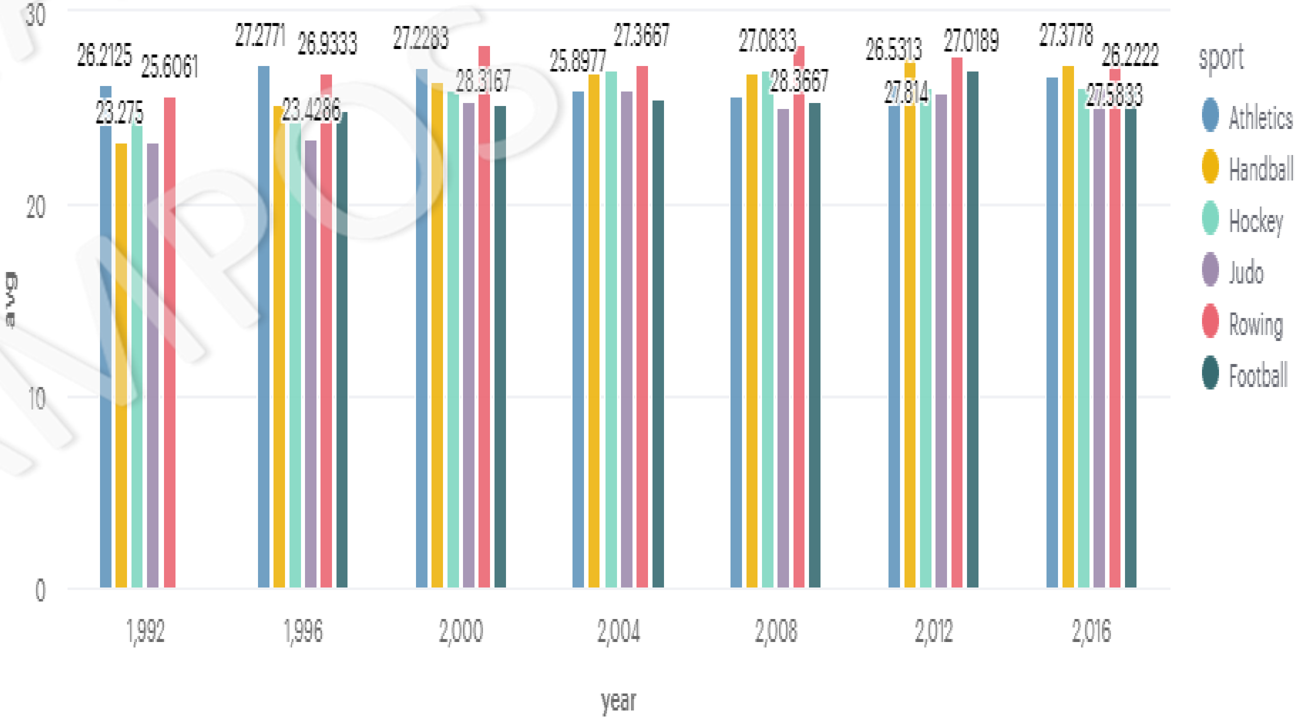
Combine the most popular sport among female athletes, average age and quantities of athletes competing in the specified sport the Olympic games.

Data from 1990 - 2016





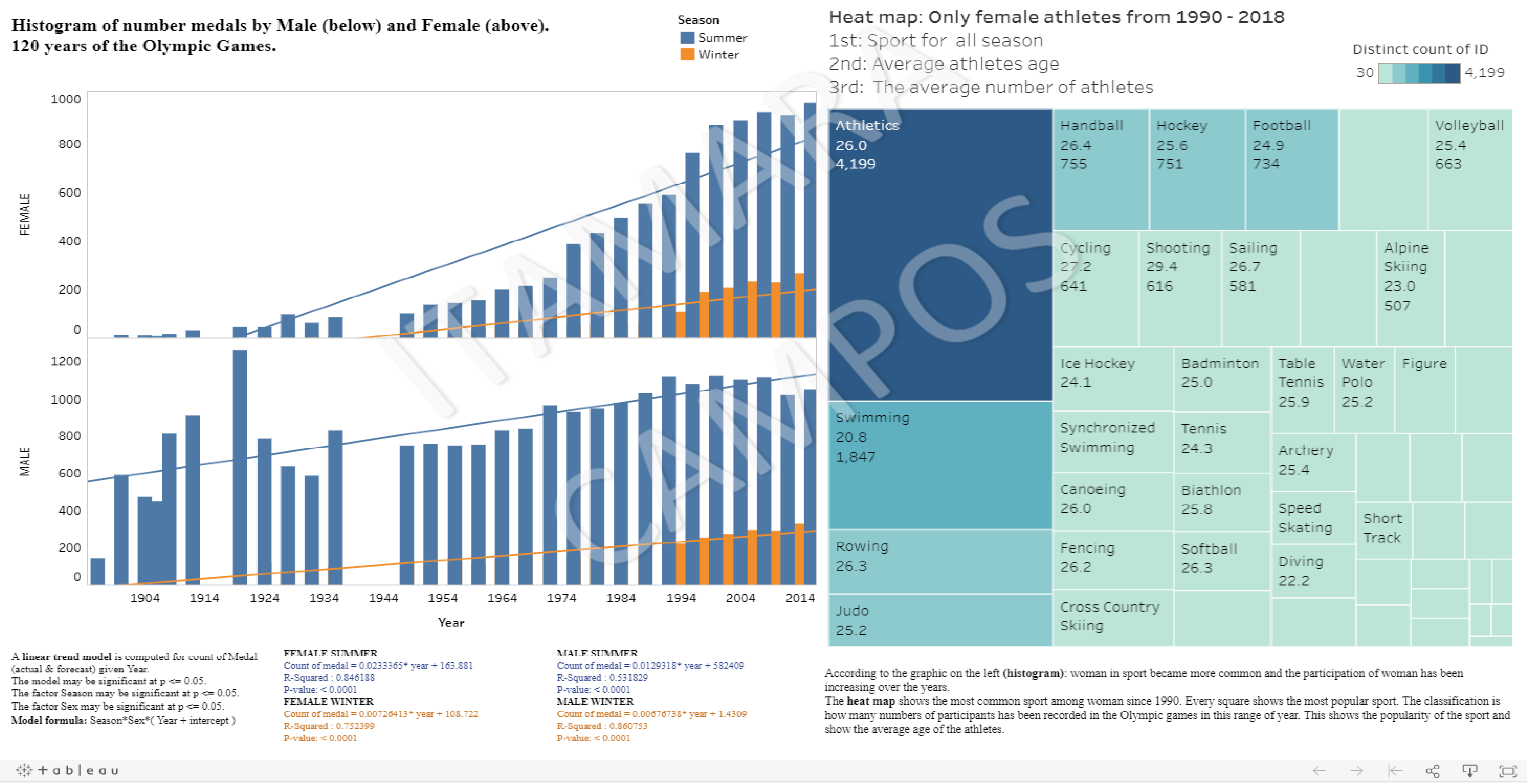
```
 SELECT COUNT(sport) as Total, year,  sport, AVG(age),  sex
FROM
 itamaracampos.noc_regions inner JOIN itamaracampos.athlete_events_copy
 on itamaracampos.noc_regions.noc = itamaracampos.athlete_events_copy.noc
Where Medal is not NULL and year > '1990' and sex = 'F' and sport in ('Athletics', 'wimming', 'Rowing', 'Judo', 'Handball', 'Hockey', 'Football')
GROUP by sport,  year, sex
ORDER by Total DESC
```



The **heat map** shows the most common sport among woman since 1990. Every square shows the most popular sport. The classification is how many numbers of participants has been recorded in the Olympic games in this range of year. This shows the popularity of the sport and show the average age of the athletes.

The **bar map** shows the average age of the athletes for the most common sport among woman

Final visualization in Public Tableau



Histogram of number medals by Male (below) and Female (above). 120 years of the Olympic Games.

Heat map: Only female athletes from 1990 - 2018
1st: Sport for all season
2nd: Average athletes age
3rd: The average number of athletes

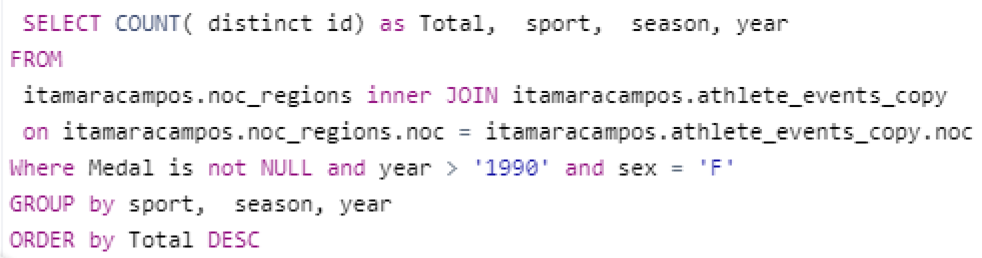
A linear trend model is computed for count of Medal (actual & forecast) given Year.
The model may be significant at p <= 0.05.
The factor Season may be significant at p <= 0.05.
The factor Sex may be significant at p <= 0.05.
Model formula: Season*Sex*(Year + intercept)

FEMALE SUMMER
Count of medal = 0.0233365* year + 163.881
R-Squared : 0.846188
P-value: < 0.0001
FEMALE WINTER
Count of medal = 0.00726413* year + 108.722
R-Squared : 0.752399
P-value: < 0.0001

MALE SUMMER
Count of medal = 0.0129318* year + 582409
R-Squared : 0.531829
P-value: < 0.0001
MALE WINTER
Count of medal = 0.00676738* year + 1.4309
R-Squared : 0.860753
P-value: < 0.0001

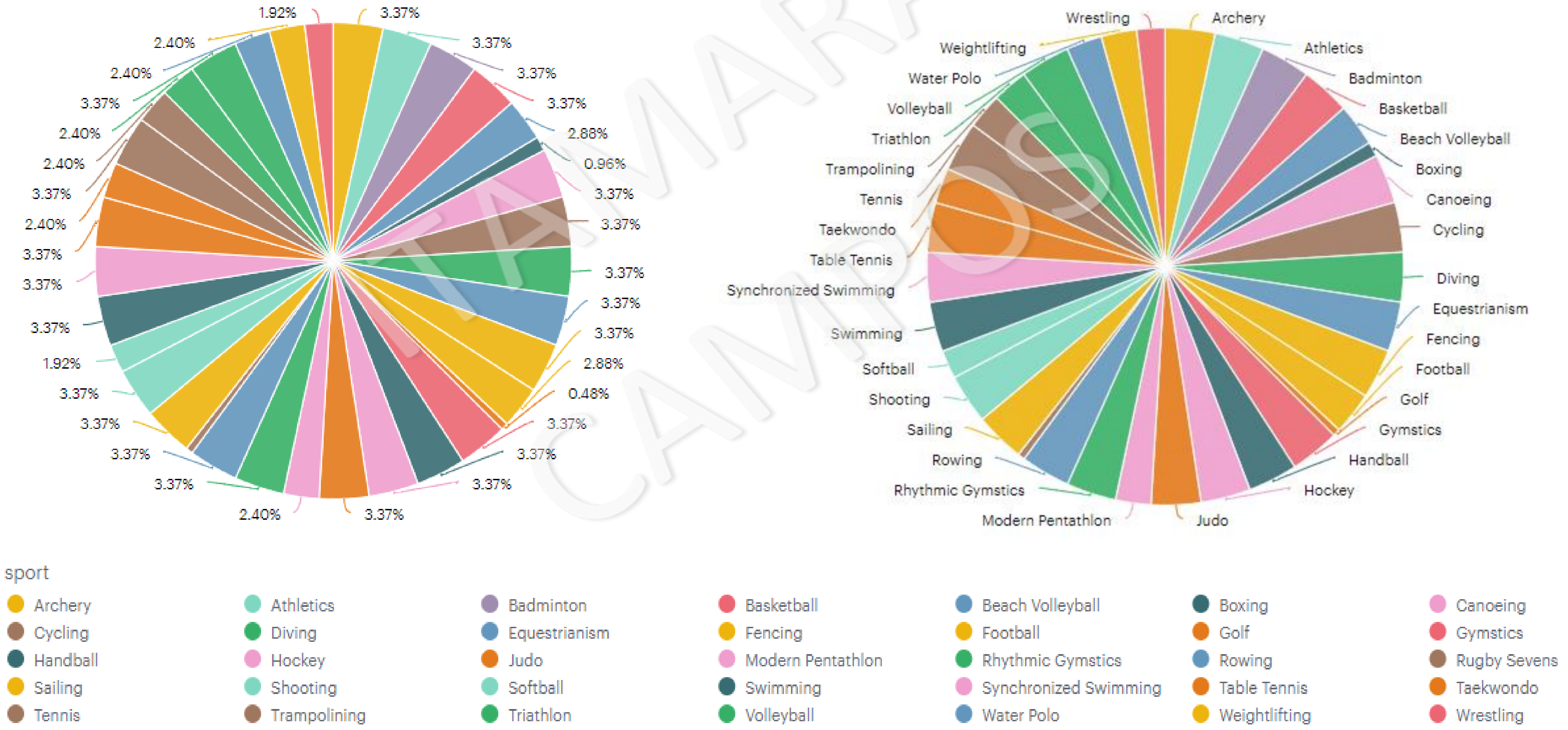
According to the graphic on the left **(histogram)**: woman in sport became more common and the participation of woman has been increasing over the years.
The **heat map** shows the most common sport among woman since 1990. Every square shows the most popular sport. The classification is how many numbers of participants has been recorded in the Olympic games in this range of year. This shows the popularity of the sport and show the average age of the athletes.

Additional Graphics build on *app.mode* using the same SQL code

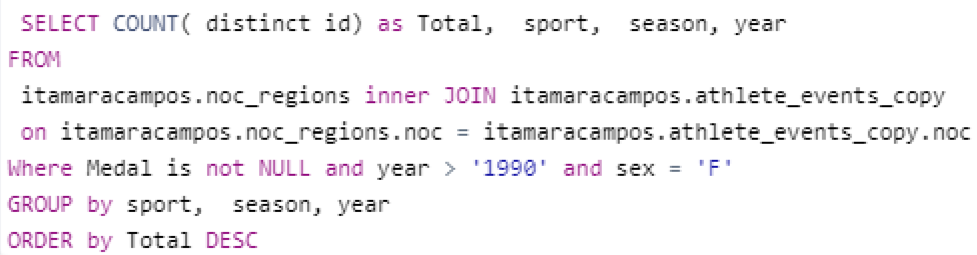


```
 SELECT COUNT( distinct id) as Total,  sport,  season, year
FROM
 itamaracampos.noc_regions inner JOIN itamaracampos.athlete_events_copy
 on itamaracampos.noc_regions.noc = itamaracampos.athlete_events_copy.noc
Where Medal is not NULL and year > '1990' and sex = 'F'
GROUP by sport,  season, year
ORDER by Total DESC
```

Filter by summer season: Pizza graphic

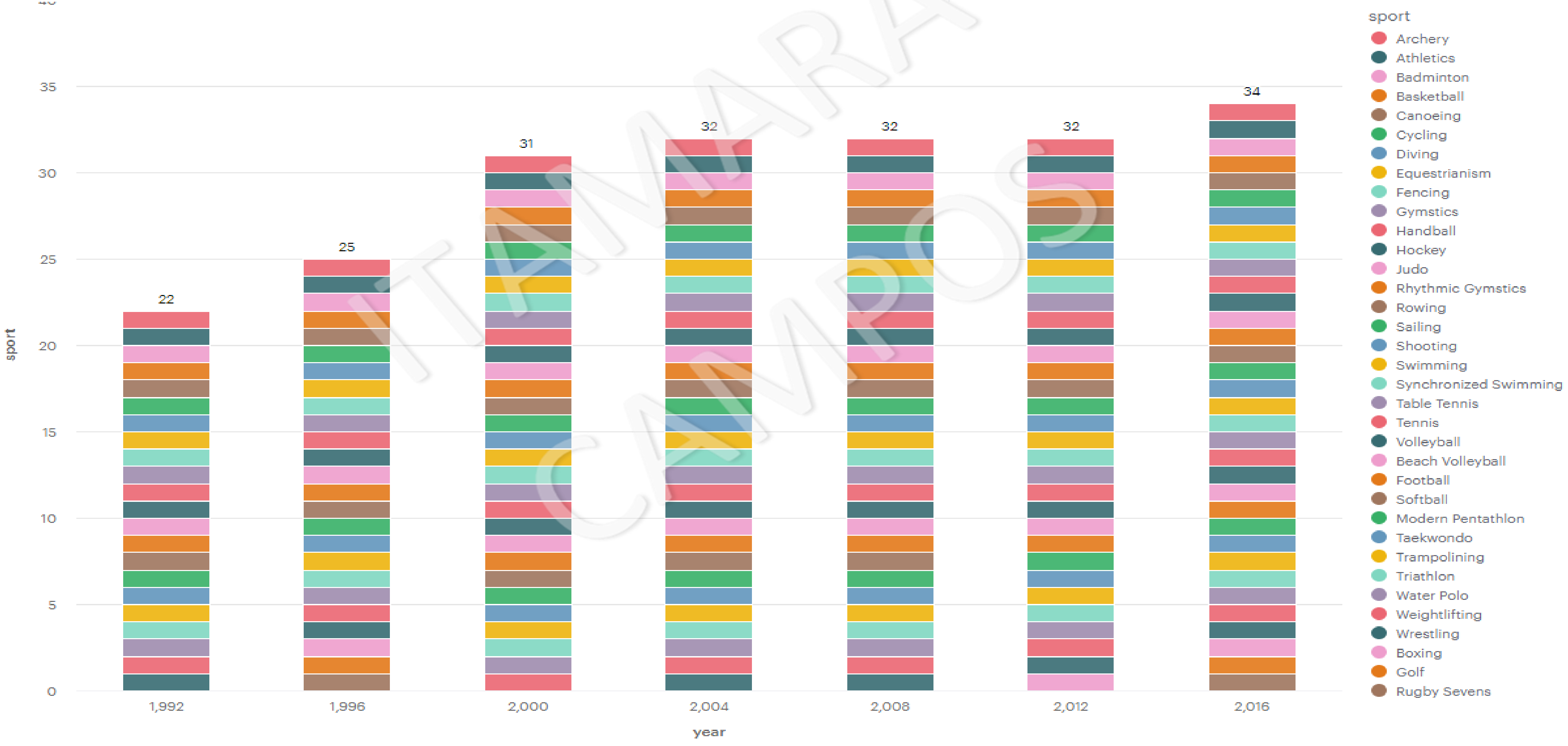


Additional Graphics build on *app.mode* using the same SQL code

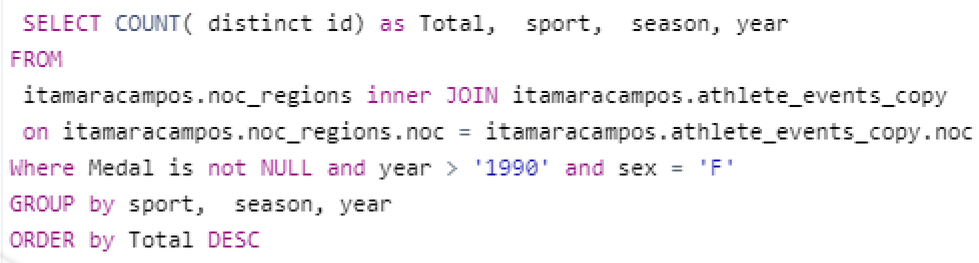


```
 SELECT COUNT( distinct id) as Total,  sport,  season, year
FROM
 itamaracampos.noc_regions inner JOIN itamaracampos.athlete_events_copy
 on itamaracampos.noc_regions.noc = itamaracampos.athlete_events_copy.noc
Where Medal is not NULL and year > '1990' and sex = 'F'
GROUP by sport,  season, year
ORDER by Total DESC
```

Filter by summer season: Bar graphic

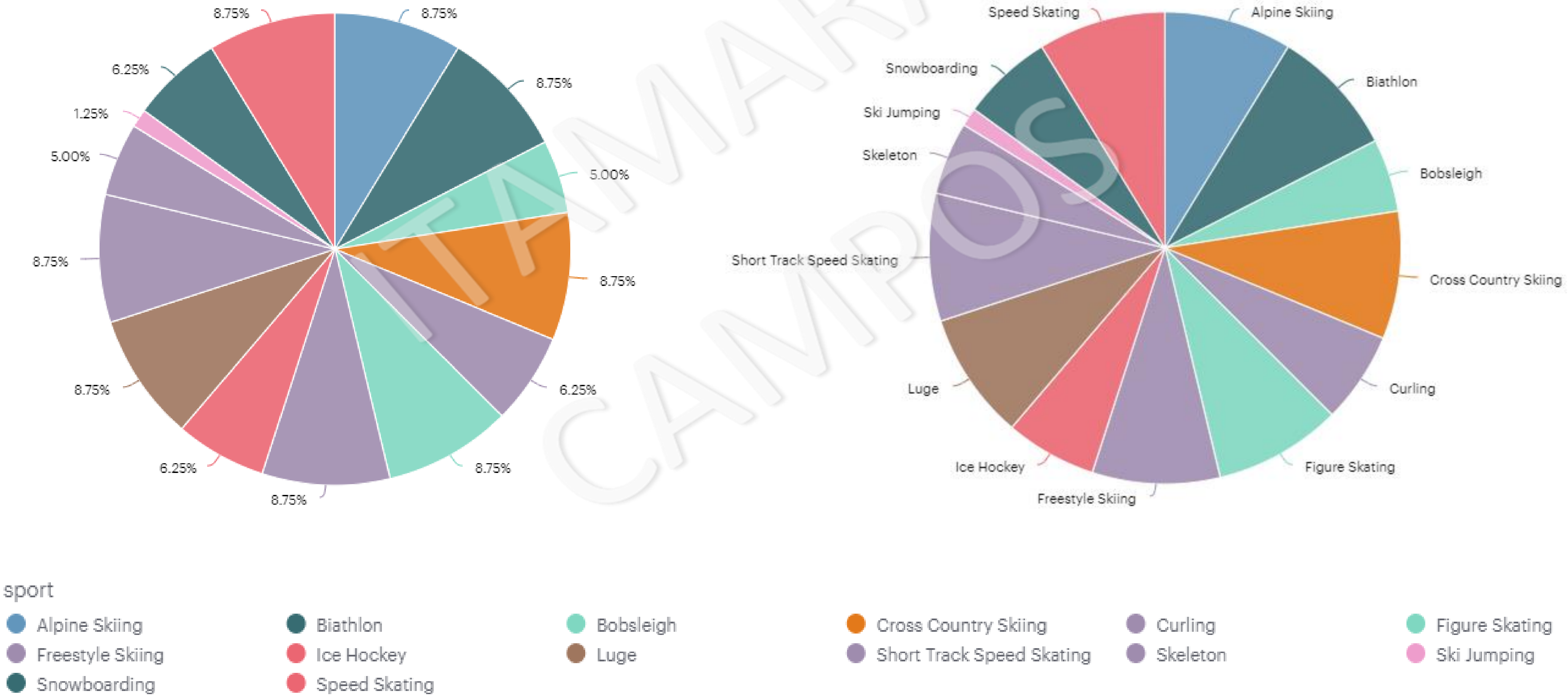


Additional Graphics build on *app.mode* using the same SQL code

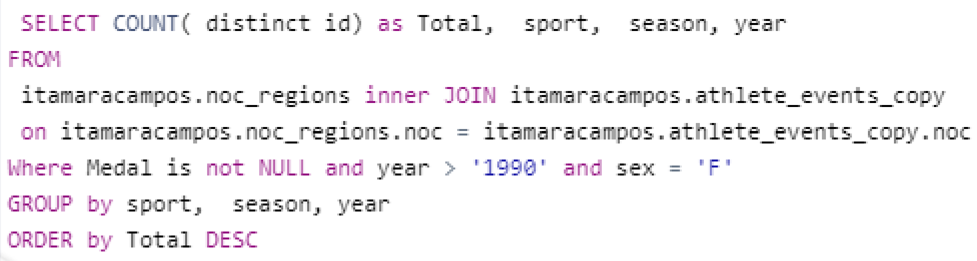


```
 SELECT COUNT( distinct id) as Total,  sport,  season, year
FROM
 itamaracampos.noc_regions inner JOIN itamaracampos.athlete_events_copy
 on itamaracampos.noc_regions.noc = itamaracampos.athlete_events_copy.noc
Where Medal is not NULL and year > '1990' and sex = 'F'
GROUP by sport,  season, year
ORDER by Total DESC
```

Filter by winter season: Pizza graphic

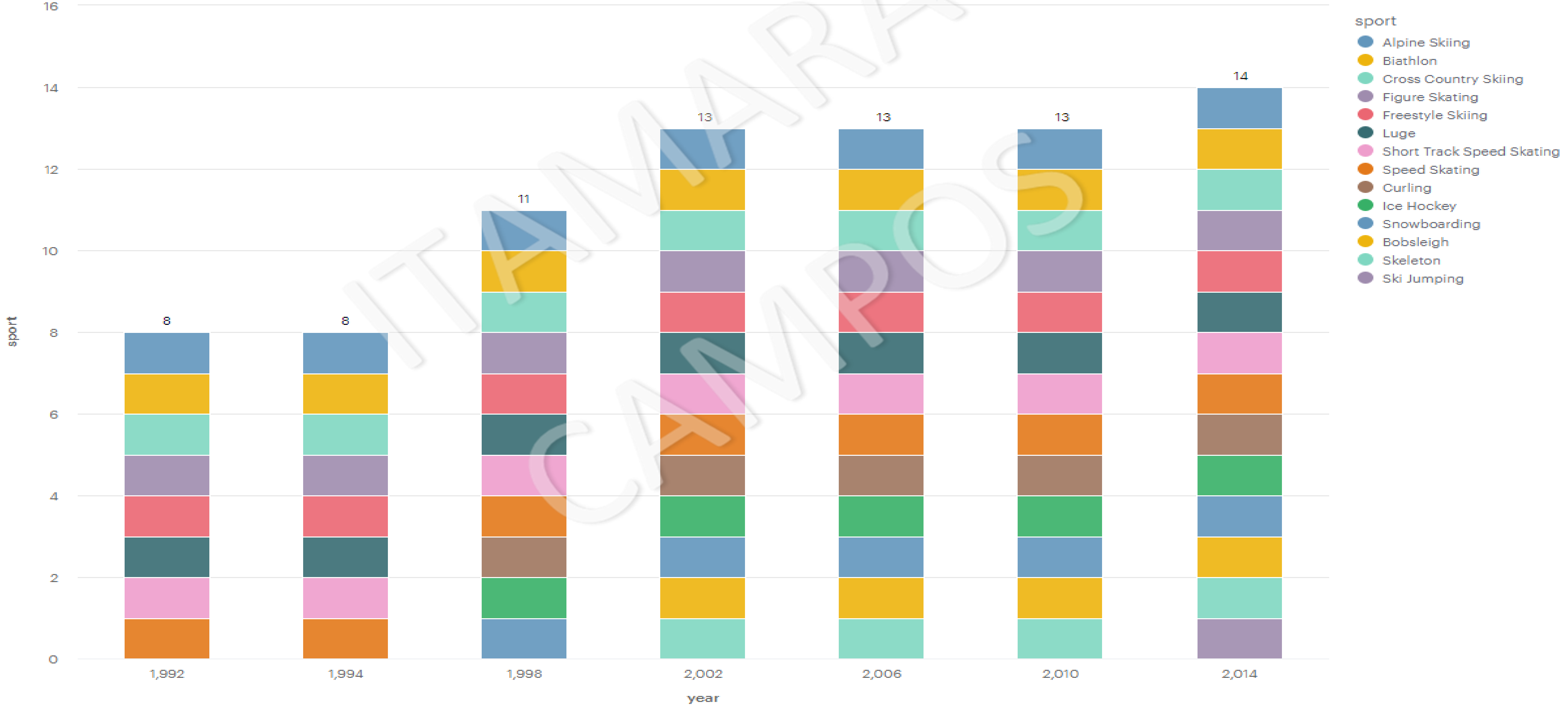


Additional Graphics build on *app.mode* using the same SQL code

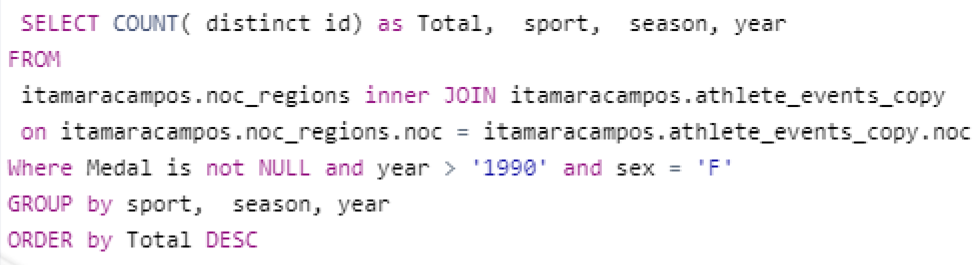


```
 SELECT COUNT( distinct id) as Total,  sport,  season, year
FROM
 itamaracampos.noc_regions inner JOIN itamaracampos.athlete_events_copy
 on itamaracampos.noc_regions.noc = itamaracampos.athlete_events_copy.noc
Where Medal is not NULL and year > '1990' and sex = 'F'
GROUP by sport,  season, year
ORDER by Total DESC
```

Filter by winter season: Bar graphic



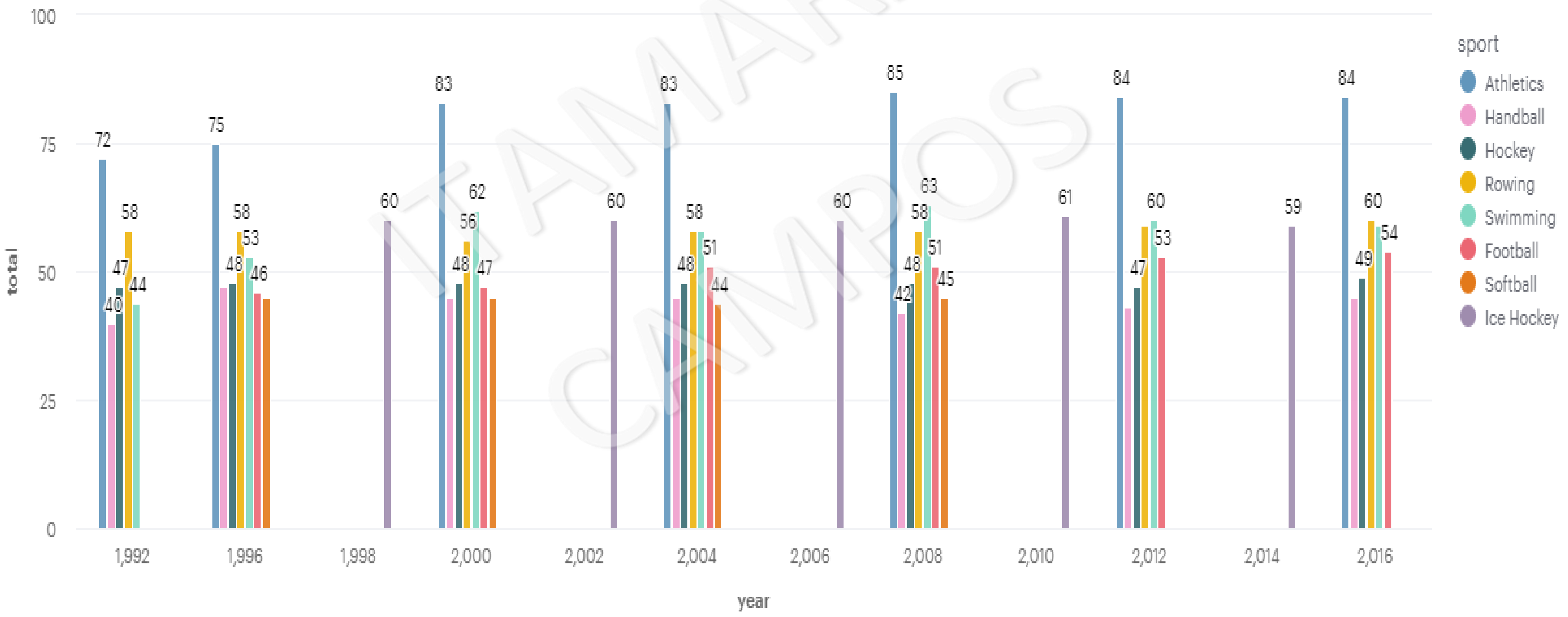
Additional Graphics build on *app.mode* using the same SQL code



```sql
 SELECT COUNT( distinct id) as Total,  sport,  season, year
FROM
 itamaracampos.noc_regions inner JOIN itamaracampos.athlete_events_copy
 on itamaracampos.noc_regions.noc = itamaracampos.athlete_events_copy.noc
Where Medal is not NULL and year > '1990' and sex = 'F'
GROUP by sport,  season, year
ORDER by Total DESC
```

Filter by *Total* variable in the SQL code in the right: Bar graphic

Show the most common sport by female athletes distribute over the years



Conclusion

Confirmation of all hypotheses.

1. Women participated more in the sports, even more then man, that also include gain some medals.
2. Gender equality increased in various sports.
3. New sports (modalities) in the in the Olympic games after 1990

More into the data

✓ Olympic games every 2 years, after 1990. Alternate season (winter and summer)

✓ Average of woman age in the games are 25 years old. But all ages has record in the database.

✓ Popular sport amount woman after 1990: Athletics, Handball, Hockey, Rowing, Swimming, Football, Softball, Ice Hockey

✓ New modalities has been inserted in the Olympic Games woman overs the years (non uniform):

 ➢ Summer, from 1992 to 2016: 12 new modalities

 ➢ Winter, from 1992 to 2016: 6 new modalities

Recommendations and Actions

This data and information must be having some importance for

 ➢ Sport clothes → Invest in comfortable accessories for the most common sport, …

 ➢ Sport equipment → Adapt to all ages, easy to move, gym location, ..

 ➢ Electronic stores → TV, smartphone… sales can increase in the OG events.

 ➢ School → invest in physical activities for the most popular games at school

 ➢ Nutritional sport store → What the athletes eat to make better performance? Can be sell to no-athletes?

More….

❑ Since Olympic games is increasing amount woman. Maybe create more facilities only for woman to practice more sport?

❑ Gym specialize in Olympic games, as a leisure for all ages?

❑ Correlate this data with heath organization data: How old the people that practice sport can live? What about the medical historic?

❑ How heath they can eat to be able to make a different between have heath lifestyle or/and a lifestyle of athlete?

Mode.app for SQL code and some graphics

https://app.mode.com/editor/sql_specializatio/reports/923bf69db30d/queries/1de06d6f6796

Tableau public some graphics

Sports_OG_v01 | Tableau Public

Sports_OG_v02 | Tableau Public

Sports_OG_v03 | Tableau Public