Olympics games

SQL for Data Science Capstone Project

Oslo, Norway

Itamara campos

Olympics games

- **Milestone week 01**

Milestone week 01

Develop Project Proposal

Sports became very popular nowadays for many reasons and in the variety of genders and ages. Looking the database of athletes over years it is possible to see an increased number of data inside of the database, which indicate the over the years this topic gets more and more data. The reason varies but is too soon the point out why the interest in sports has been increase, might be for pure entertainment or for heath reason or both.

1) Why sport became so popular?
2) Or which sport became so popular?
3) Who might be interest in this data?
4) What is the age of my target audience after the analysis of this data?

Hypothesis

1. I believe over the years women participate more in the sports, that also include gain some medals.
2. Gender equality (men and women) must be increased in various sports, but I can't tell which one.
3. New sports (modalities) in the Olympic games. Maybe an option?

Approach

- Sport, year, sex, country or city
- Medals overs years by woman; Medals overs years by man
- Which sport has more medals by men and woman
- Age, medals and years.

Choose data set

Client 3. SportsStarts (Olympics data set).

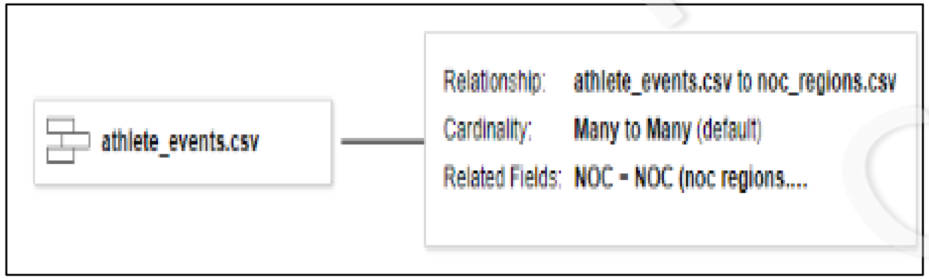
Because I like sport, and I can relate myself as a future client. When came to buy clothes, chose brands,…

The relationship between them is using the key-attribute: **NOC variable**

•The table ***athlete_events*** has a lot categories that include **NOC** attribute that connects the table ***noc_regions***

•I believe the ERD type is **One to Many**

Screenshot from the relation create in Tableau

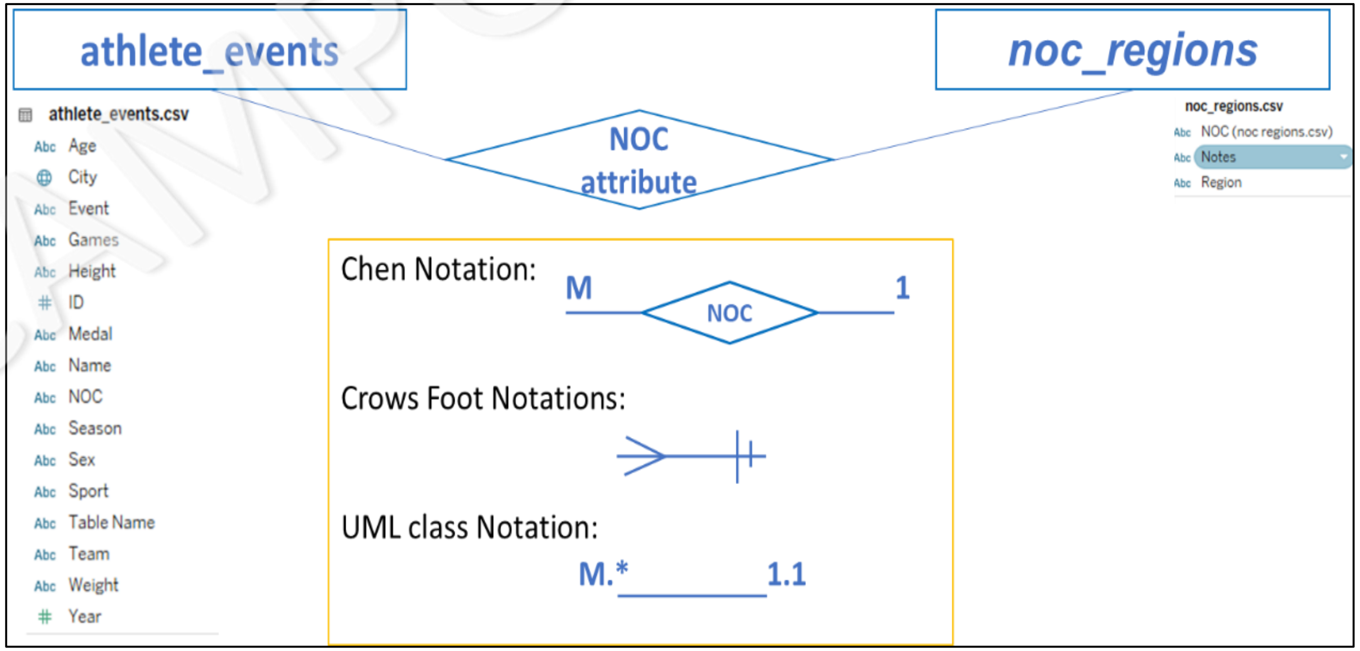

athlete_events.csv

Relationship: athlete_events.csv to noc_regions.csv

Cardinality: Many to Many (default)

Related Fields: NOC = NOC (noc regions....

ERD Diagram and relation



Olympics games

- **Milestone week 02**

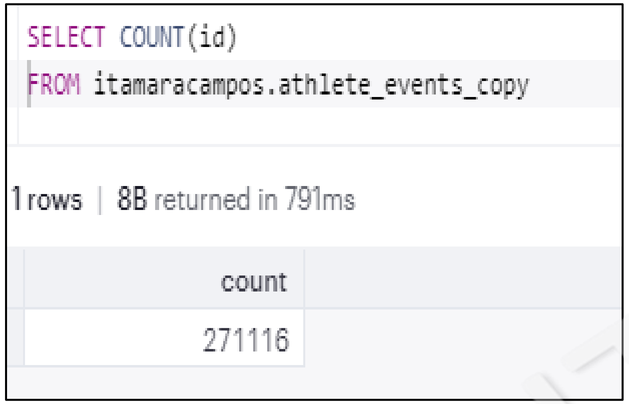
Milestone week 02

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)

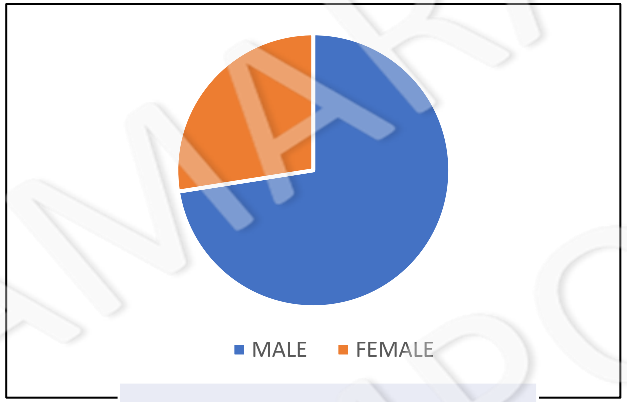
After a look how many has ***medals.*** I believe medal it will be a way the describe success of the statistics for all genders (Pizza right)

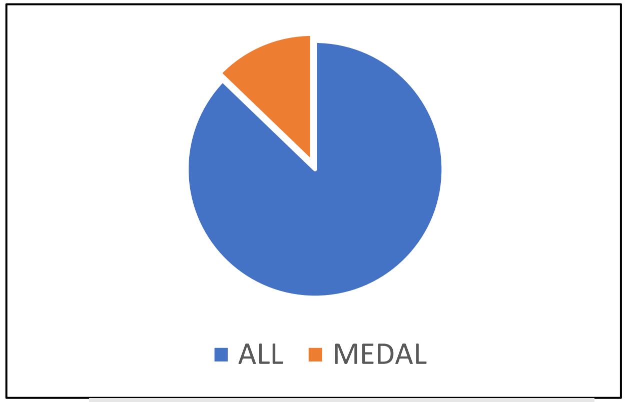


```
SELECT COUNT(id)
FROM itamaracampos.athlete_events_copy
```

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

  count
 271116
```





PORCENTAGERS	
72.51287272 %	27.48712728 %
MAN	FEMALE

14.67379% gain medals

Looking the percentages

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

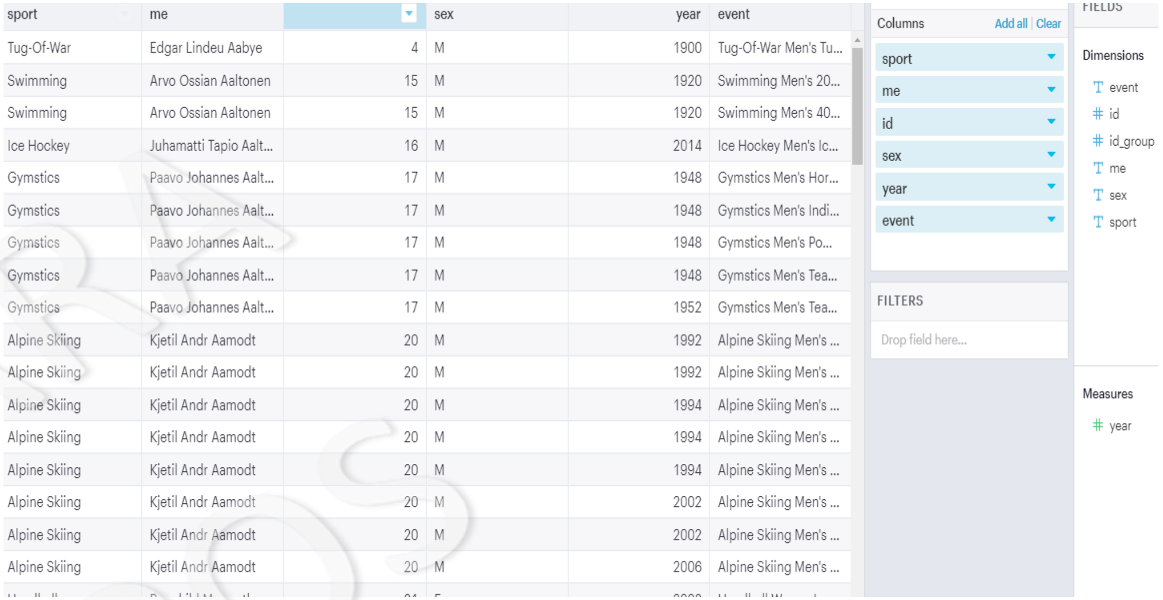
2. The gender equality (men and woman) must be increased in various sports, but I can’t tell which one. (but I believe there are more women in the events overall)

   I still can’t not confirm which sport became more popular between woman… I need more investigation

Milestone week 02

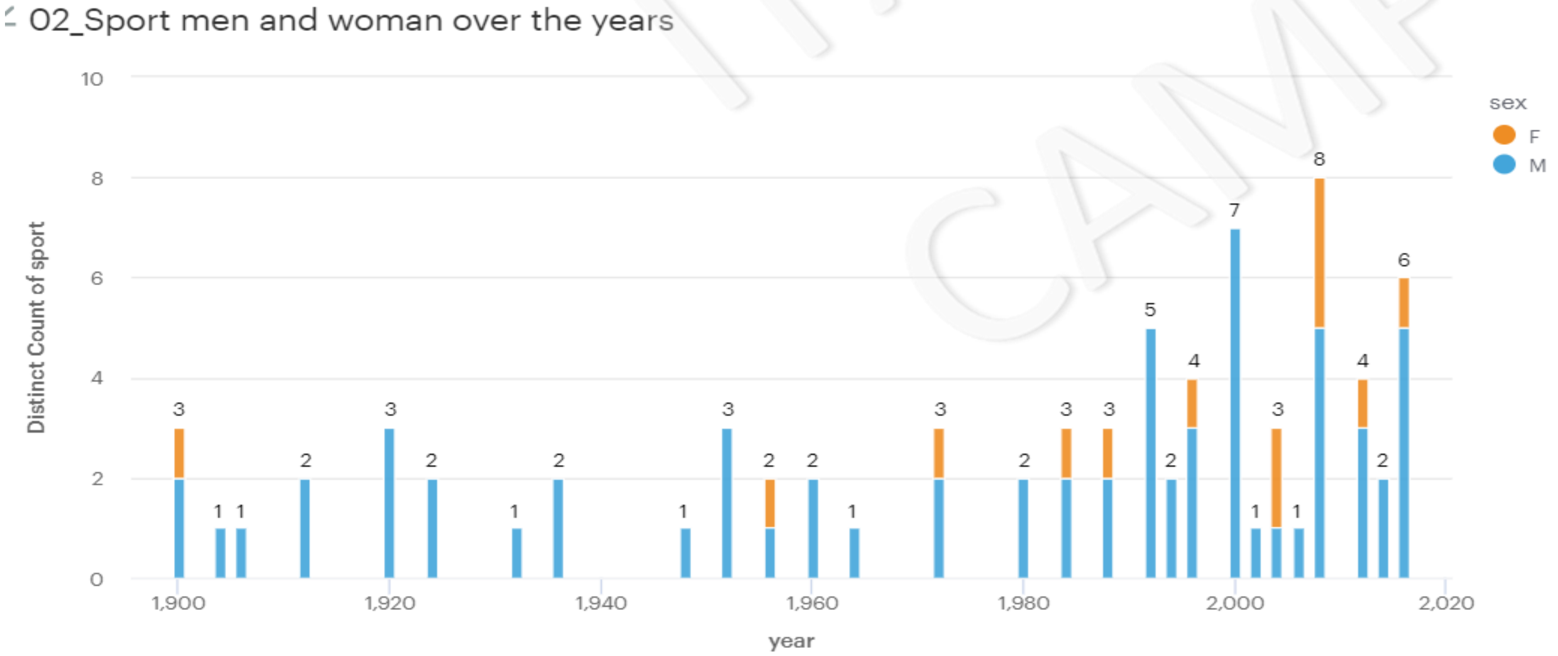
I believe over the years more people participated of the Olympiads game.
Only consider where the fields medal is not NULL

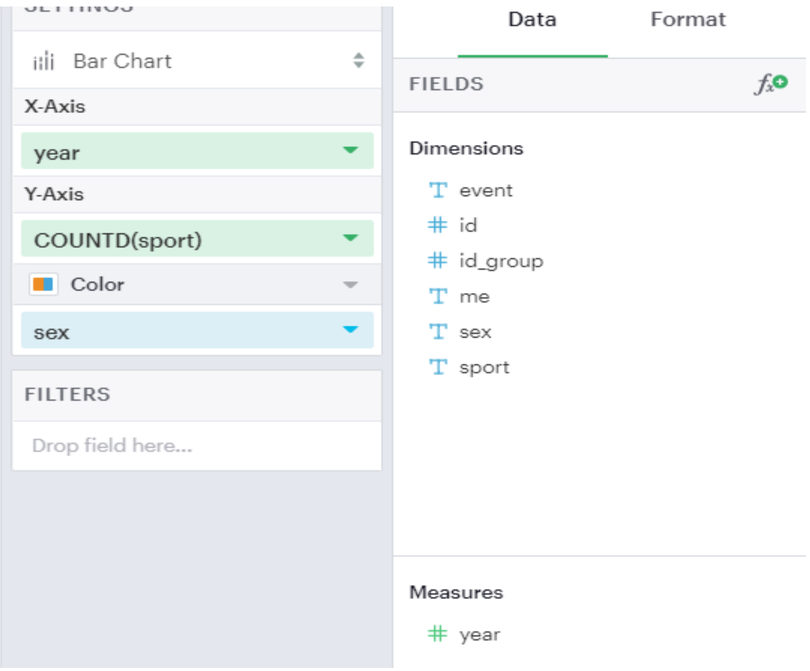
```
SELECT
   COUNT( distinct id) as id_group,
   me, id, sex, year, sport, event
FROM itamaracampos.athlete_events_copy
WHERE medal is not NULL
GROUP by id, me, sex, year, sport, event
```



sport	me		sex	year	event
Tug-Of-War	Edgar Lindeu Aabye	4	M	1900	Tug-Of-War Men's Tu...
Swimming	Arvo Ossian Aaltonen	15	M	1920	Swimming Men's 20...
Swimming	Arvo Ossian Aaltonen	15	M	1920	Swimming Men's 40...
Ice Hockey	Juhamatti Tapio Aalt...	16	M	2014	Ice Hockey Men's Ic...
Gymstics	Paavo Johannes Aalt...	17	M	1948	Gymstics Men's Hor...
Gymstics	Paavo Johannes Aalt...	17	M	1948	Gymstics Men's Indi...
Gymstics	Paavo Johannes Aalt...	17	M	1948	Gymstics Men's Po...
Gymstics	Paavo Johannes Aalt...	17	M	1948	Gymstics Men's Tea...
Gymstics	Paavo Johannes Aalt...	17	M	1952	Gymstics Men's Tea...
Alpine Skiing	Kjetil Andr Aamodt	20	M	1992	Alpine Skiing Men's ...
Alpine Skiing	Kjetil Andr Aamodt	20	M	1992	Alpine Skiing Men's ...
Alpine Skiing	Kjetil Andr Aamodt	20	M	1994	Alpine Skiing Men's ...
Alpine Skiing	Kjetil Andr Aamodt	20	M	1994	Alpine Skiing Men's ...
Alpine Skiing	Kjetil Andr Aamodt	20	M	1994	Alpine Skiing Men's ...
Alpine Skiing	Kjetil Andr Aamodt	20	M	2002	Alpine Skiing Men's ...
Alpine Skiing	Kjetil Andr Aamodt	20	M	2002	Alpine Skiing Men's ...
Alpine Skiing	Kjetil Andr Aamodt	20	M	2006	Alpine Skiing Men's ...

Columns: sport, me, id, sex, year, event





Bar Chart

X-Axis

year

Y-Axis

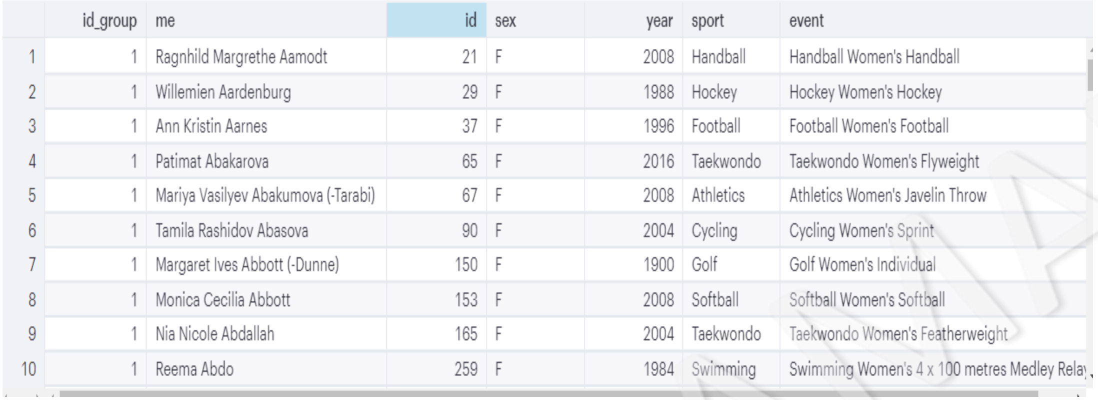
COUNTD(sport)

Color

sex

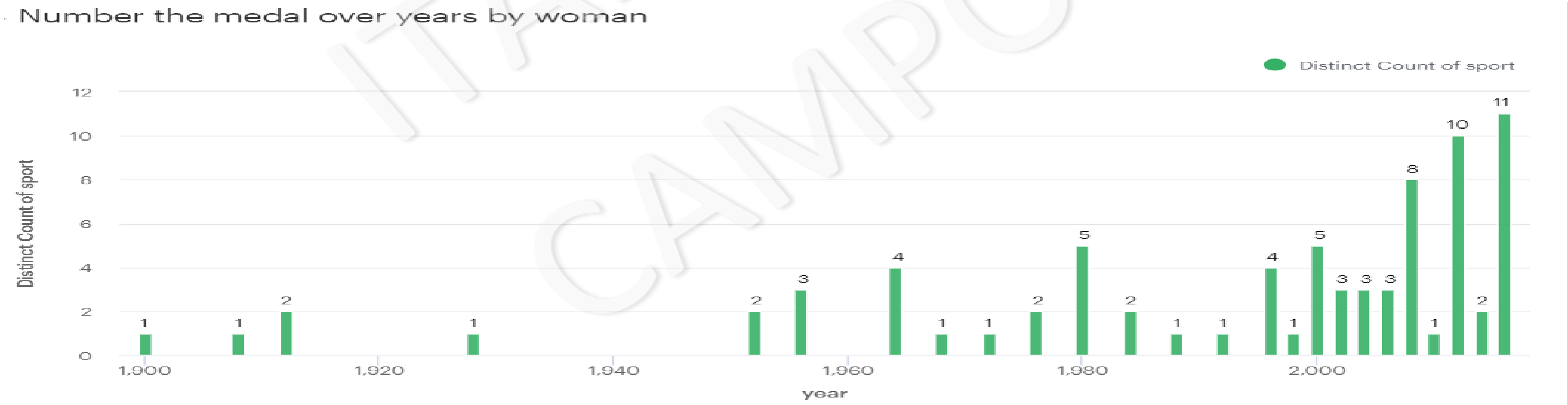
Milestone week 02

3 : Only consider where the fields medal is not NULL only Sex F



	id_group	me	id	sex	year	sport	event
1	1	Ragnhild Margrethe Aamodt	21	F	2008	Handball	Handball Women's Handball
2	1	Willemien Aardenburg	29	F	1988	Hockey	Hockey Women's Hockey
3	1	Ann Kristin Aarnes	37	F	1996	Football	Football Women's Football
4	1	Patimat Abakarova	65	F	2016	Taekwondo	Taekwondo Women's Flyweight
5	1	Mariya Vasilyev Abakumova (-Tarabi)	67	F	2008	Athletics	Athletics Women's Javelin Throw
6	1	Tamila Rashidov Abasova	90	F	2004	Cycling	Cycling Women's Sprint
7	1	Margaret Ives Abbott (-Dunne)	150	F	1900	Golf	Golf Women's Individual
8	1	Monica Cecilia Abbott	153	F	2008	Softball	Softball Women's Softball
9	1	Nia Nicole Abdallah	165	F	2004	Taekwondo	Taekwondo Women's Featherweight
10	1	Reema Abdo	259	F	1984	Swimming	Swimming Women's 4 x 100 metres Medley Relay

```
SELECT
   COUNT( distinct id) as id_group,
   me, id, sex, year, sport, event
FROM itamaracampos.athlete_events_copy
WHERE medal is not NULL and sex = 'F’
GROUP by id, me, sex, year, sport, event
```



Confirm of the hypotheses:

Yes, the graphic show that over the years women became, ate least, more participate of the Olympics games.

Milestone week 02

Initial Hypothesis:

1. I believe 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, but I can’t tell which one.
3. New sports (modalities) in the Olympic games. Maybe an option?

Hypothesis changed plan.

1. I believe 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, ~~but I can’t tell which one~~. ✓
3. New sports (modalities) in the Olympic games. Maybe an option? Need investigation.

What additional questions are you seeking to answer?

Who might be interest in this data?

What is the age of my target audience after the analysis of this data?

Olympics games

- **Milestone week 03**

Milestone week 03

Dive Deeper

Look deeper into the features you are investigating, consider:

1. Relationships / Correlation, Pearson Correlation
2. Linear Regression for future prediction (if the relationship is linear)
3. Textual Analysis for TF-IDF (Term Frequency-Inverse Document Frequency; Row-based and column-based, stop-word removal?

Specify 1-2 correlations you discovered. List the fields that you found to be correlated and describe what you learned from these correlations.

Go Broader

Expand the features you are investigating. Look for connections/relationships that you may have initially missed.

1. What jumps out at you now?
2. Use the descriptive stats to point you to features that you may now want to consider.

What key terms did you discover in any text analysis, for whom? Any themes? If you are not analyzing text, summarize what other things you are considering in your analysis?

New Metric

Create 1 or 2 new metrics to track relationships of data you discovered. Explain why you created them.

Milestone week 03. Analyzes

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.

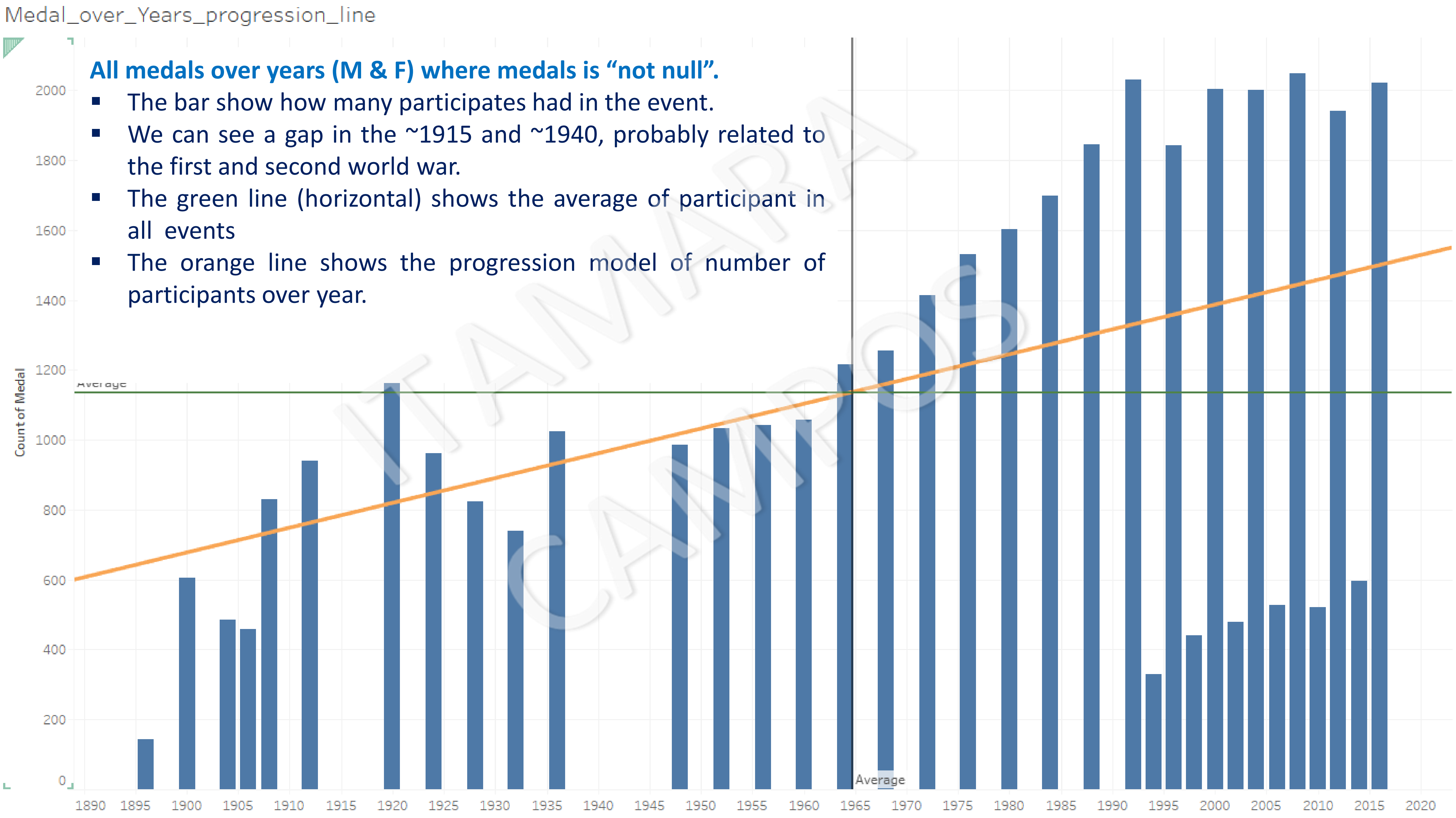
P-value

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

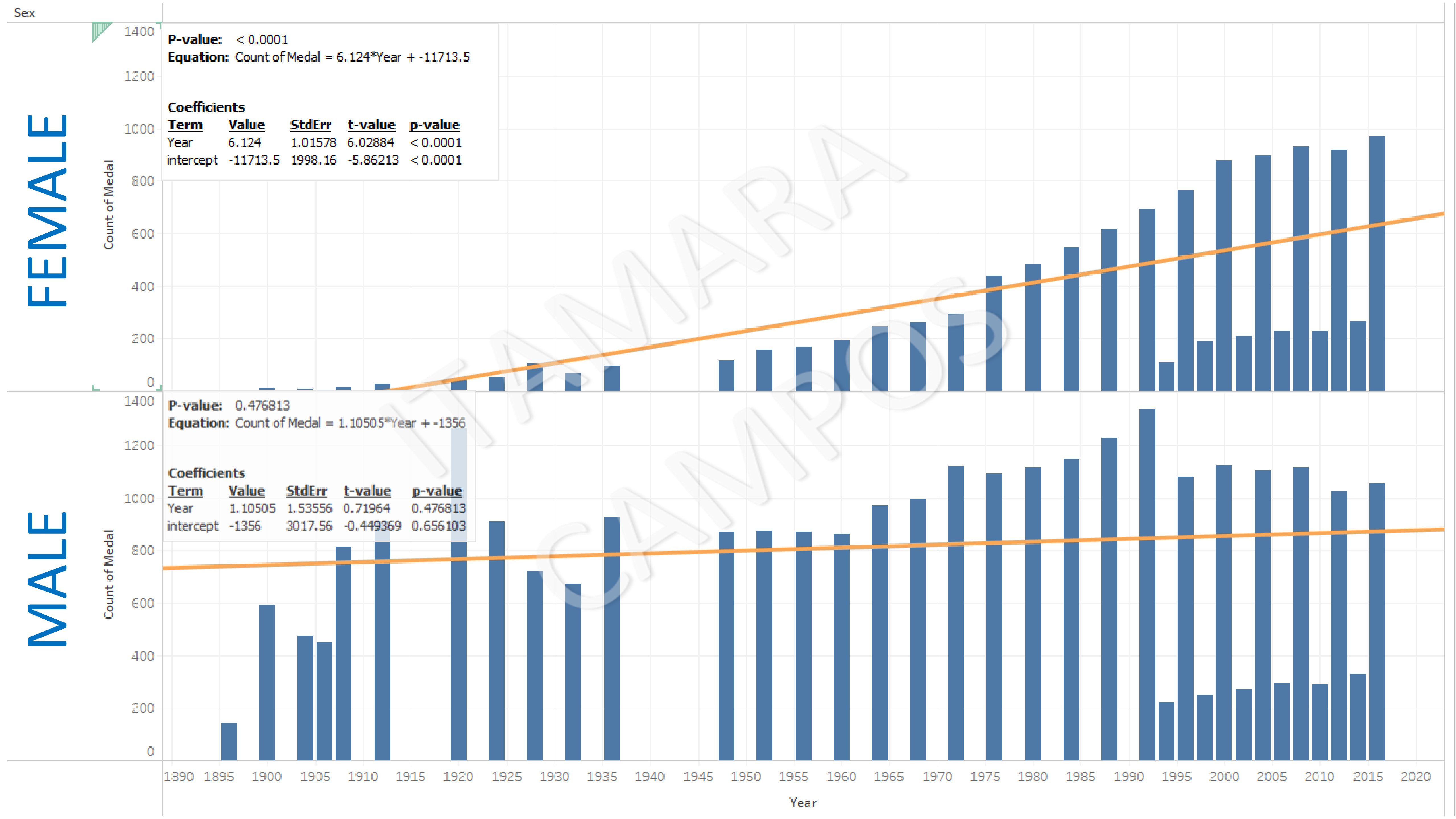
The R-squared is also an important measure when assessing if the model is suitable and tells us whether the model effectively fits our data. The R-squared is measured on a scale from 0-1; the closer to 1 the more effective the model.

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



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.



P-value: < 0.0001
Equation: Count of Medal = 6.124*Year + -11713.5

Coefficients

Term	Value	StdErr	t-value	p-value
Year	6.124	1.01578	6.02884	< 0.0001
intercept	-11713.5	1998.16	-5.86213	< 0.0001

P-value: 0.476813
Equation: Count of Medal = 1.10505*Year + -1356

Coefficients

Term	Value	StdErr	t-value	p-value
Year	1.10505	1.53556	0.71964	0.476813
intercept	-1356	3017.56	-0.449369	0.656103

Trend Lines Model
A linear trend model is computed for count of Medal given Year. The model may be significant at p <= 0.05. The factor Sex may be significant at p <= 0.05.

Model formula: Sex*(Year + intercept)
Number of modeled observations: 69
Number of filtered observations: 0
Model degrees of freedom: 4
Residual degrees of freedom (DF): 65
SSE (sum squared error): 5.5592e+06
MSE (mean squared error): 85526.2
R-Squared: 0.513129
Standard error: 292.449
p-value (significance): < 0.0001
Analysis of Variance:

Field	DF	SSE	MSE	F	p-value
Sex	2	4776534.2	2.38827e+06	27.9244	< 0.0001

Individual trend lines:

Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
F	Year	< 0.0001	32	Year	6.124	1.01578	6.02884	< 0.0001
				intercept	-11713.5	1998.16	-5.86213	< 0.0001
M	Year	0.476813	33	Year	1.10505	1.53556	0.71964	0.476813
				intercept	-1356	3017.56	-0.449369	0.656103

The total show the formula:

Count of Medal = 7.09266* Year + 12798.6
R-Squared: 0.223416
P-value: 0.0041402

Men:
Count of Medal = 1.10505* Year + 1356
R-Squared: 0.0154509
P-value: 0.476813

Women:
Count of Medal = 6.124* Year + 11713.5
R-Squared: 0.5318
P-value < 0.0001

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.

Olympics games

- **Milestone week 04**

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?

Milestone week 02

Hypothesis changed plan. From milestone week 02:

1. I believe 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, but I can't tell which one.
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 in the notebook on model.com (link in the end of the presentation)
- ❑ The graphics were generated by Tableau
- ➢ *Events* columns should be call as subcategory of the sport. That could be misleading to false conclusion.

Discuss Insights Discovered

Based on the previous slides, the analyses show an increased number of woman in the Olympic games. That prove hypothesis number 1. But, to avoid repeat information and make this presentation clearer. I will bring some finds about the data. The metric it will be show in the graphics itself using visualization.

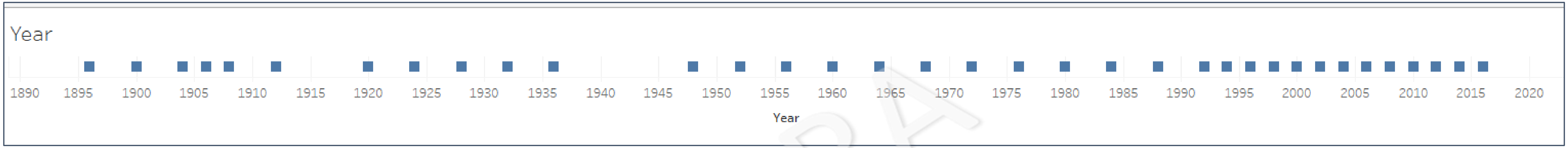
Recommendations and Actions

I will recommend:

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

Woman in the Olympics Games

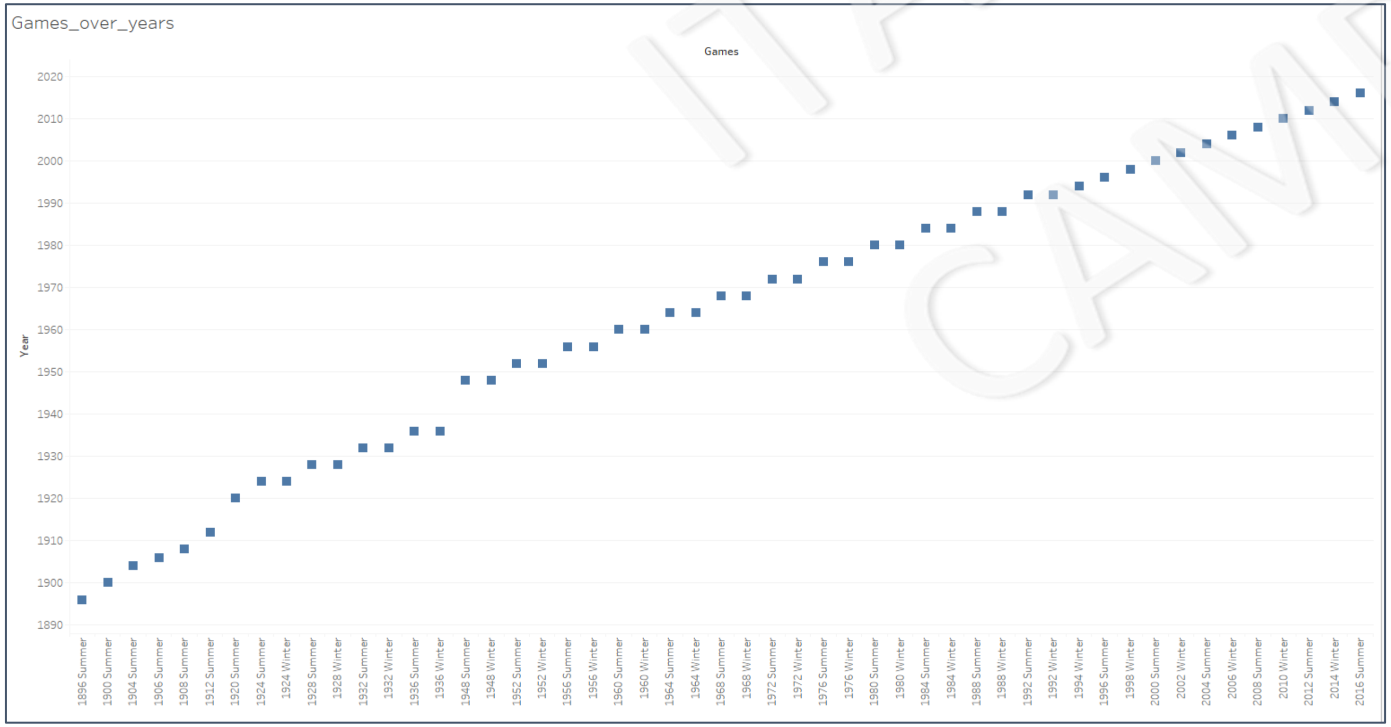
An Analise over time



Year

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.

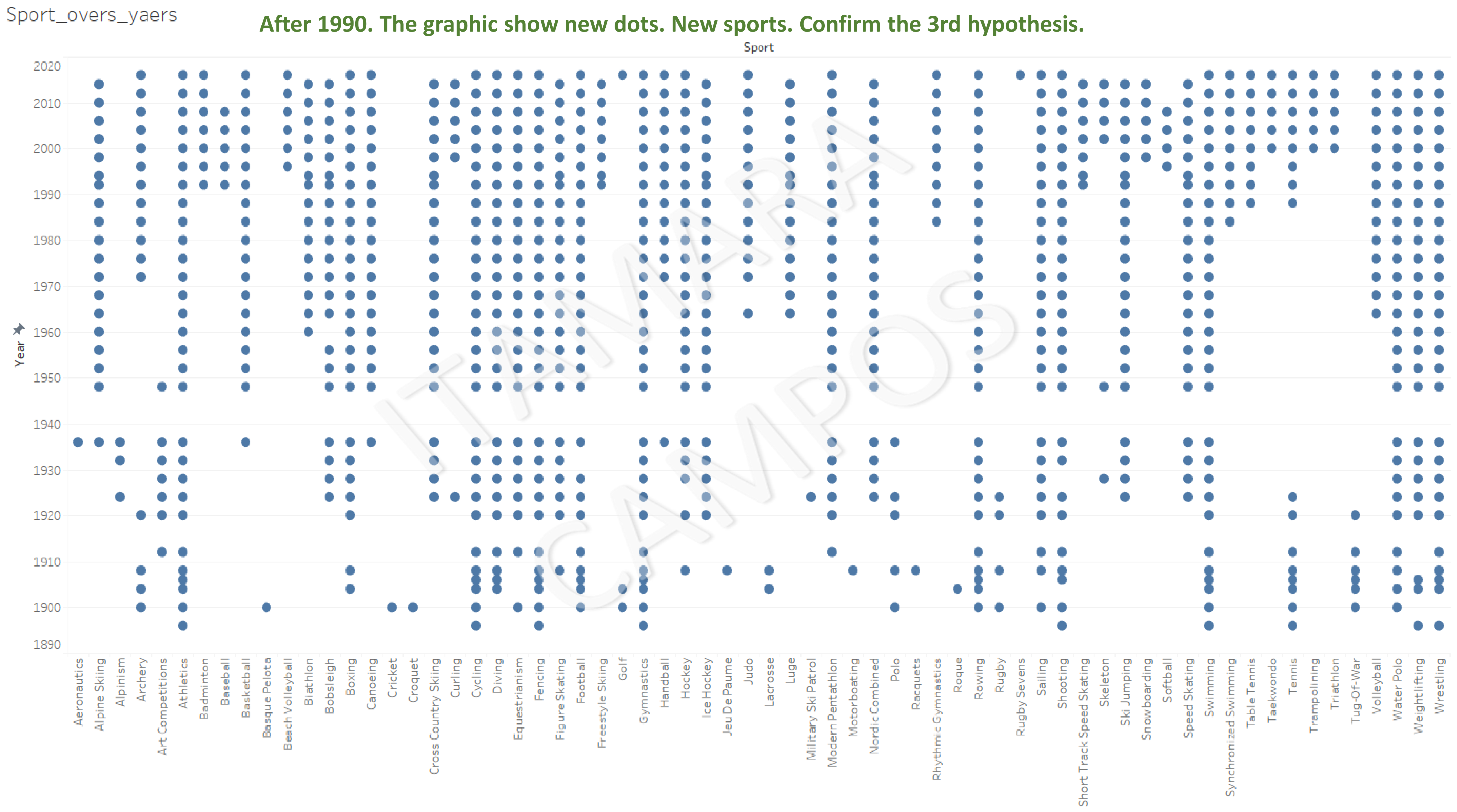


Games_over_years

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.

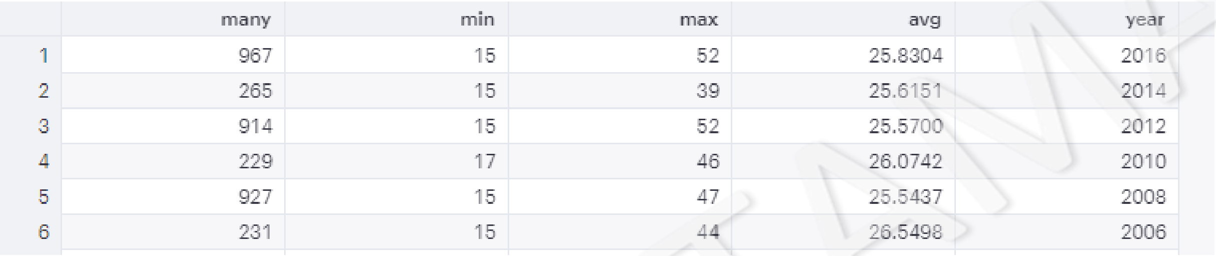
Which show more linear in the graphic after 1990.



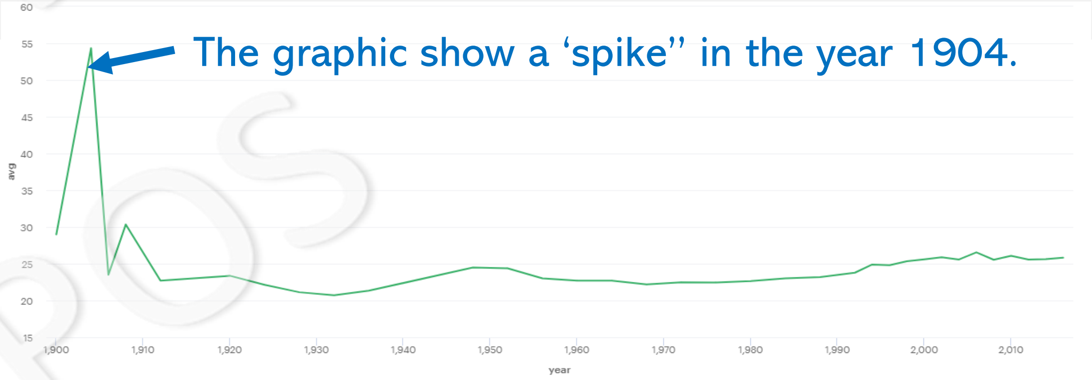
After 1990. The graphic show new dots. New sports. Confirm the 3rd hypothesis.

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
```

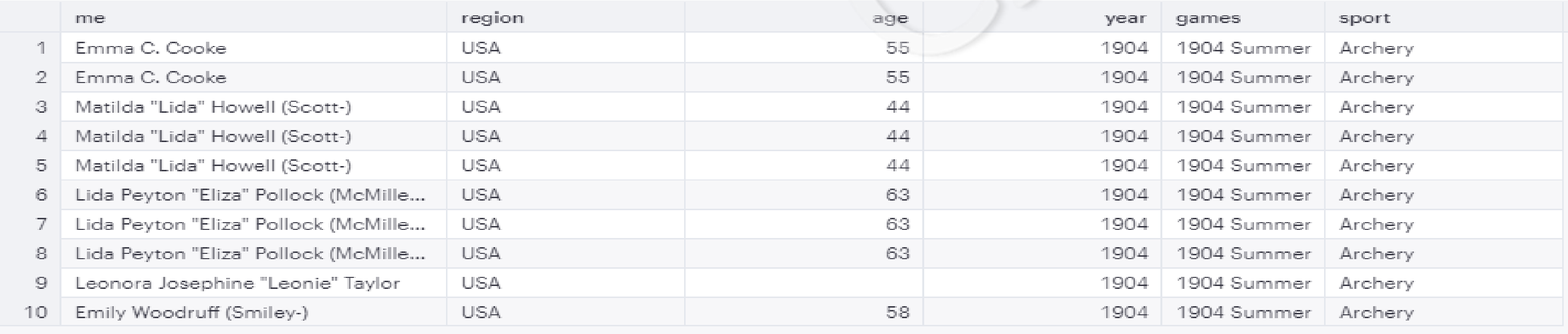


	many	min	max	avg	year
1	967	15	52	25.8304	2016
2	265	15	39	25.6151	2014
3	914	15	52	25.5700	2012
4	229	17	46	26.0742	2010
5	927	15	47	25.5437	2008
6	231	15	44	26.5498	2006



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'
```

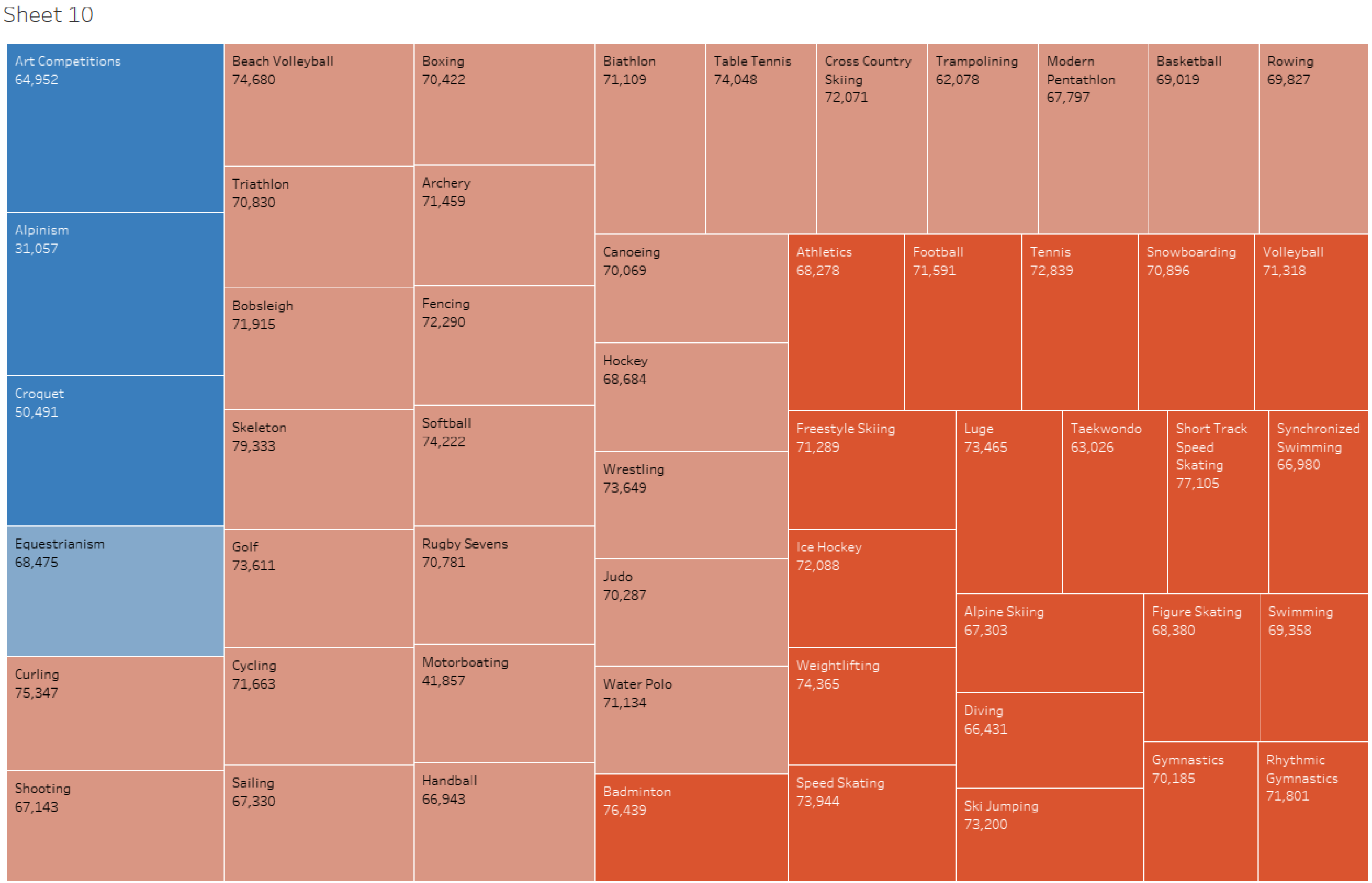


	me	region	age	year	games	sport
1	Emma C. Cooke	USA	55	1904	1904 Summer	Archery
2	Emma C. Cooke	USA	55	1904	1904 Summer	Archery
3	Matilda "Lida" Howell (Scott-)	USA	44	1904	1904 Summer	Archery
4	Matilda "Lida" Howell (Scott-)	USA	44	1904	1904 Summer	Archery
5	Matilda "Lida" Howell (Scott-)	USA	44	1904	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.

Average woman id (all participates) for all sports

Less popular sport among woman.



More popular sport among woman.

Conclusion

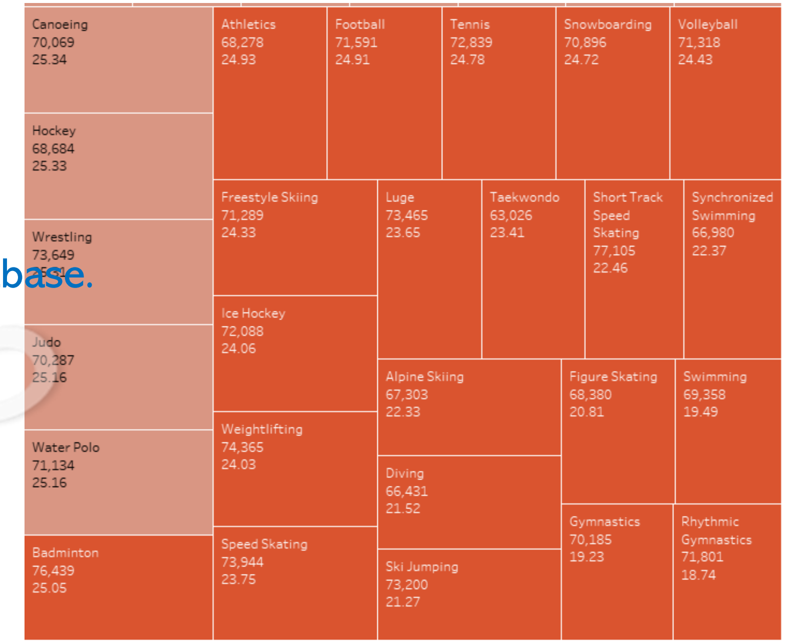
1. Women participated more in the sports, even more then man, that also include gain some medals.
2. Gender equality increased in various sports.
3. New sports (modalities) in the in the Olympic games after 1990

More into the data

Olympic games every 2 years. 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, see the graphic in the right, also with age average



Recommendations and Actions

- This data and information must be having some importance for
  - Sport clothes → Which kind sport is more popular amount woman, …
  - Sport equipment → Adapt to all ages, easy to move, comfortable, gym location…
  - Electronic stores → TV, smartphone… sales can increase in the OG events.

More….

- Since Olympic games is increasing amount woman. Maybe create a facilities for woman sport?
- Gym specialize in Olympic games, as a leisure for all ages?

Mode.app for SQL code and some graphics

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