

# Milestone 4 – SQL for Data Science Capstone Project

Juan Villaverde



# Preparing for your proposal

- Se utilizó el dataset de **SportsStats** porque, de las 3 propuestas, me pareció la más interesante, pues me gusta mucho el deporte y las olimpiadas.
- Para importar la data, se fragmento el archivo .CSV de la fuente de datos y se usó DataBricks para la visualización de los datos.
- Luego se creo una tabla donde se unen todos los archivos .CSV para tener una tabla con los datos

Cmd 2

```
1 DROP TABLE IF EXISTS athlete_events;
2
3 CREATE EXTERNAL TABLE athlete_events (
4   `ID` INT,
5   `Name` STRING,
6   `Sex` STRING,
7   `Age` INT,
8   `Height` INT,
9   `Weight` INT,
10  `Team` STRING,
11  `NOC` STRING,
12  `Games` STRING,
13  `Year` INT,
14  `Season` STRING,
15  `City` STRING,
16  `Sport` STRING,
17  `Event` STRING,
18  `Medal` STRING
19 )
20 STORED AS parquet
21 LOCATION '/tmp/athlete_events'
```

OK

Command took 0.41 seconds -- by juanvillapre2230@gmail.com at

```
1 INSERT INTO athlete_events
2 SELECT *
3 FROM athlete_events1
4 UNION ALL
5 SELECT *
6 FROM athlete_events2
7 UNION ALL
8 SELECT *
9 FROM athlete_events3
10 UNION ALL
11 SELECT *
12 FROM athlete_events4
13 UNION ALL
14 SELECT *
15 FROM athlete_events5
```

► (1) Spark Jobs

OK

Command took 14.43 seconds -- by juanvillapre2230@gmail.com at 14/11/20

Cmd 4

1 SELECT \* FROM athlete\_events

2

► (1) Spark Jobs

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year
994	559	Isvetelina Abrasheva	F	16	null	null	Bulgaria	BUL	1994 Winter	1994
995	560	Amir Abrashi	M	22	172	70	Switzerland	SUI	2012 Summer	2012
996	561	Thomas Abratis	M	26	184	73	Germany	GER	1994 Winter	1994
997	561	Thomas Abratis	M	26	184	73	Germany	GER	1994 Winter	1994
998	562	Pawe Abratkiewicz	M	21	183	84	Poland	POL	1992 Winter	1992
999	562	Pawe Abratkiewicz	M	21	183	84	Poland	POL	1992 Winter	1992
1000	562	Pawe Abratkiewicz	M	27	183	84	Poland	POL	1998 Winter	1998

Truncated results, showing first 1000 rows.



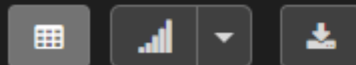
Command took 5.55 seconds -- by juanvillapre2230@gmail.com at 14/11/2021 17:29:48 on Milestone1

```
SELECT * FROM noc_regions
```

► (1) Spark Jobs

	NOC ▲	region ▲	notes ▲
1	AFG	Afghanistan	null
2	AHO	Curacao	Netherlands Antilles
3	ALB	Albania	null
4	ALG	Algeria	null
5	AND	Andorra	null
6	ANG	Angola	null
7	ANT	Antigua	Antigua and Barbuda

Showing all 230 rows.

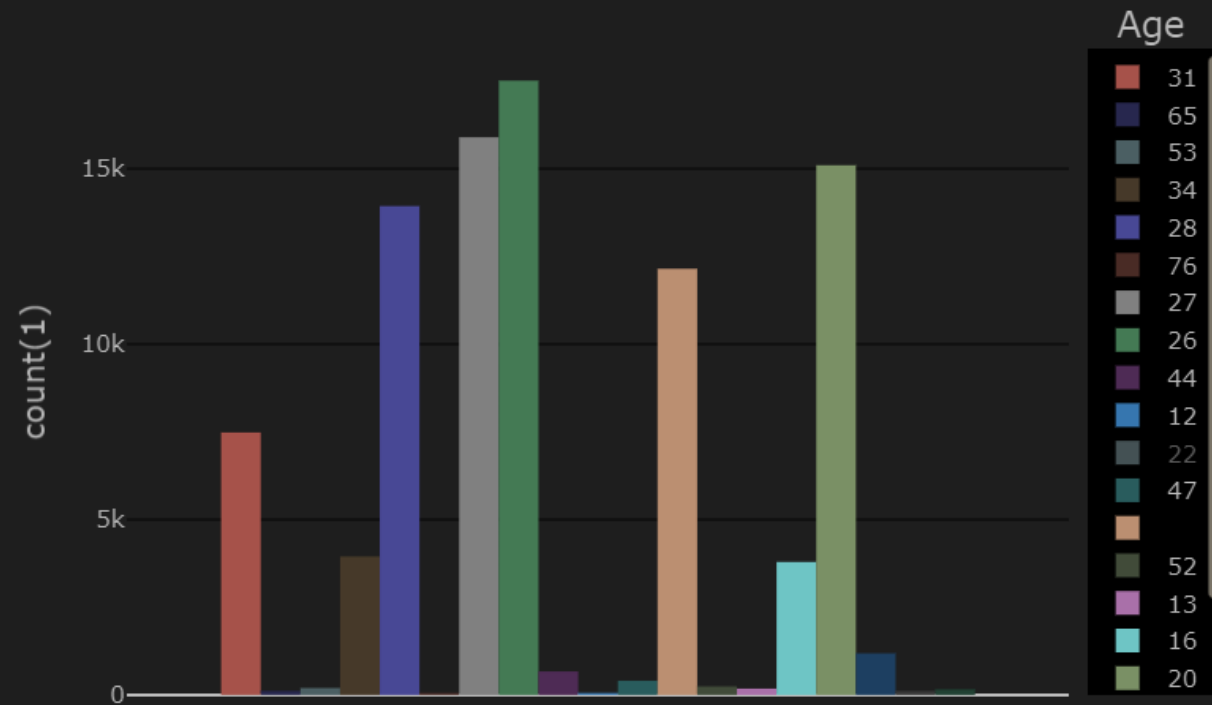


Command took 0.42 seconds -- by juanvillapre2230@gmail.com at 14/11/2021 17:33:34 on Milestone1

Cmd 6

```
SELECT Age, COUNT(*)  
FROM athlete_events  
GROUP BY Age
```

► (2) Spark Jobs

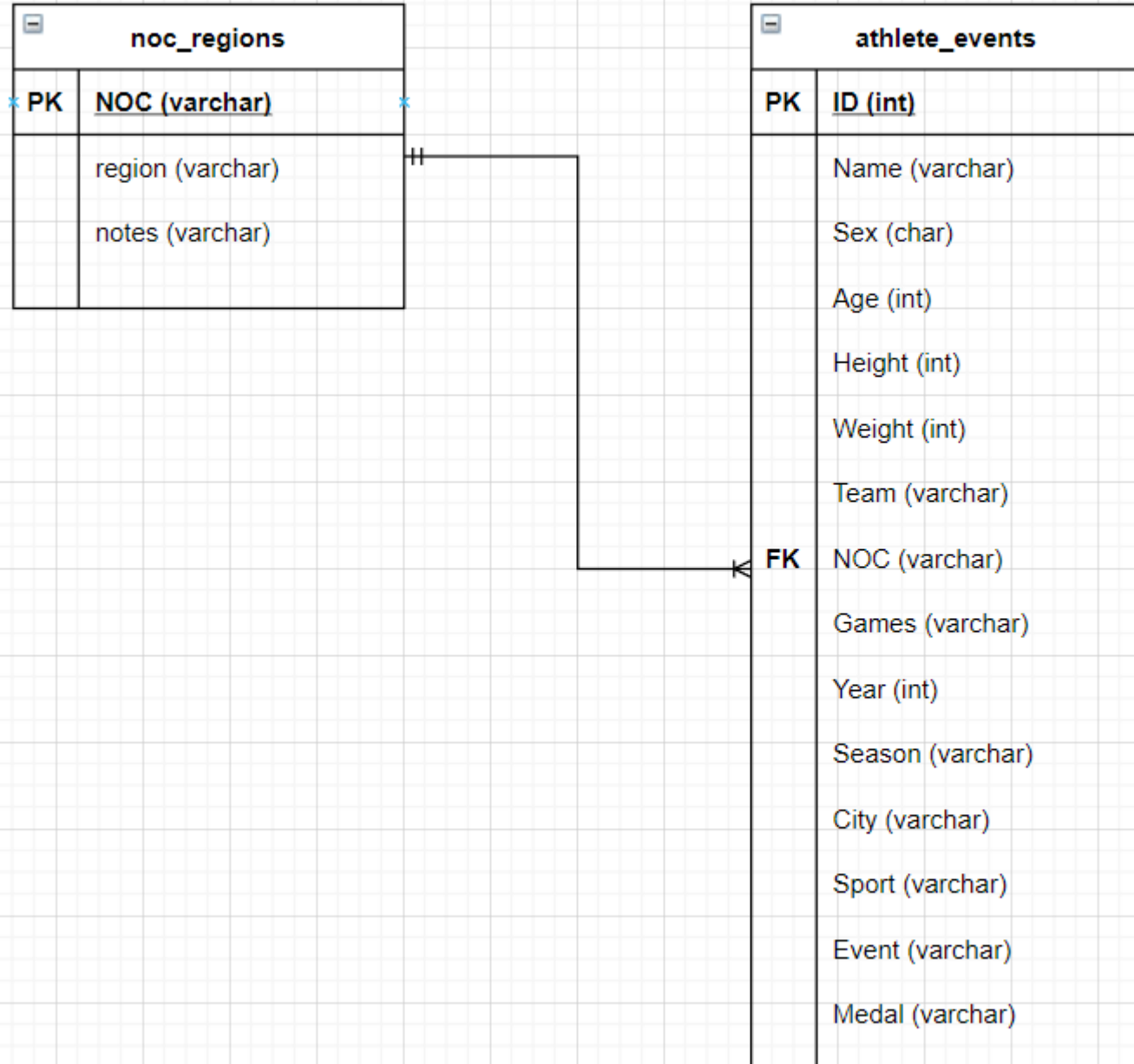


Only showing the first twenty series.

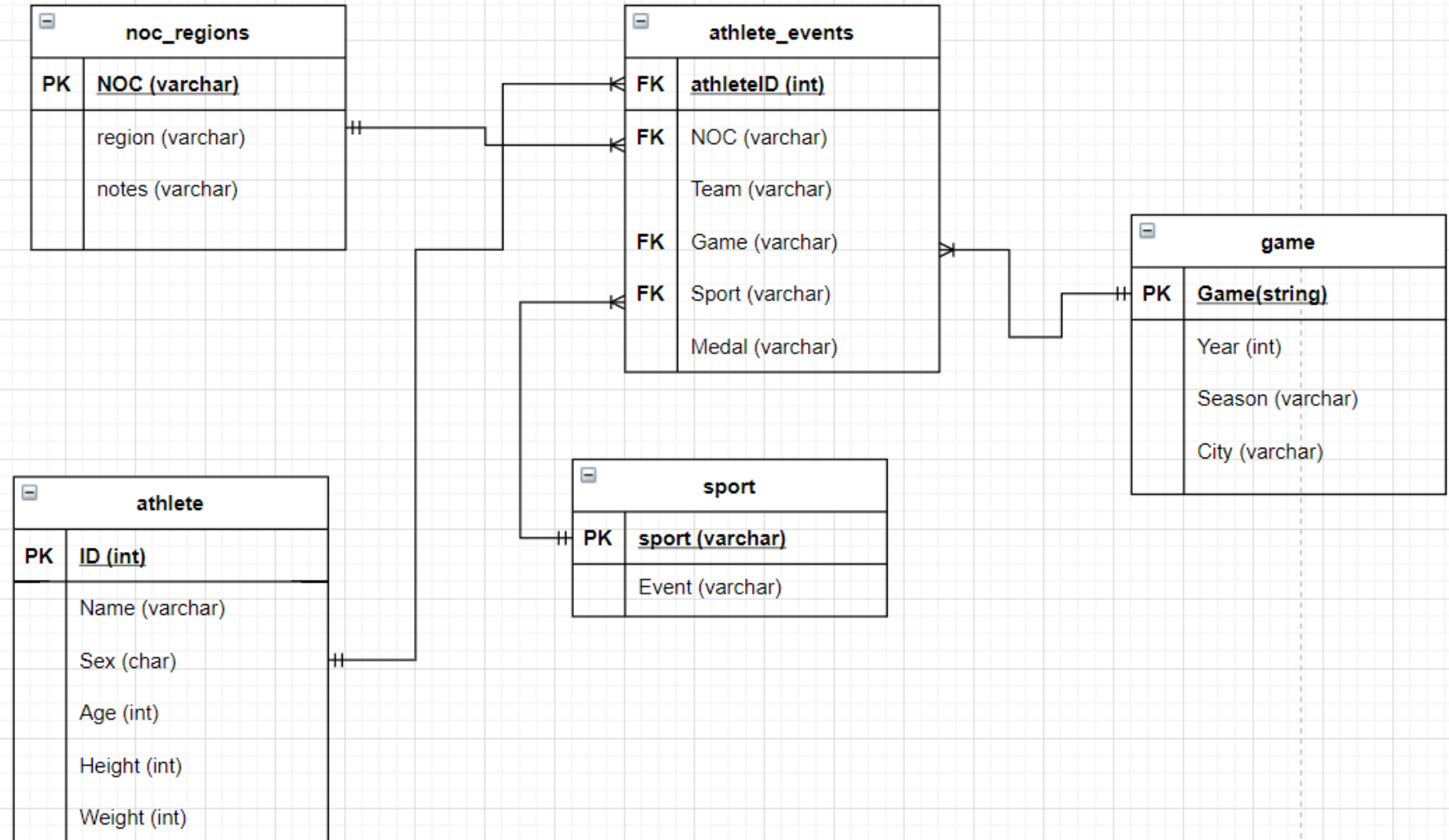


Command took 4.48 seconds -- by juanvillapre2230@gmail.com at 14/11/2021 17:35:54 on Milestone1

# OLD ERD



# NEW ERD





# Develop Project Proposal

## ➤ DESCRIPTION

- El proyecto busca analizar los atletas y los deportes que formaron parte de la historia de las olimpiadas, desde el inicio hasta la actualidad.
- Se busca analizar todos los campos posibles y existentes dentro de la data entregada por parte de la fuente.

## ➤ QUESTION

- ¿Cuál fue el primer país en tener más medallas en la primera olimpiada?
- ¿Cuál es la distribución de edades entre todos los atletas que han participado?
- ¿Cuántas medallas de oro, plata y bronce se han entregado en todas las olimpiadas?
- ¿Cuántos deportes o disciplinas distintas se han registrado en todas las olimpiadas?
- ¿Cuáles fueron las primeras disciplinas realizadas en las primeras olimpiadas realizadas?



## ➤ HYPOTHESIS

- Estados Unidos tuvo la mayor cantidad de medallas
- Muchos de los atletas participaron entre 20 y 30 años
- Más de 3 mil medallas
- Más de 60 deportes o disciplinas

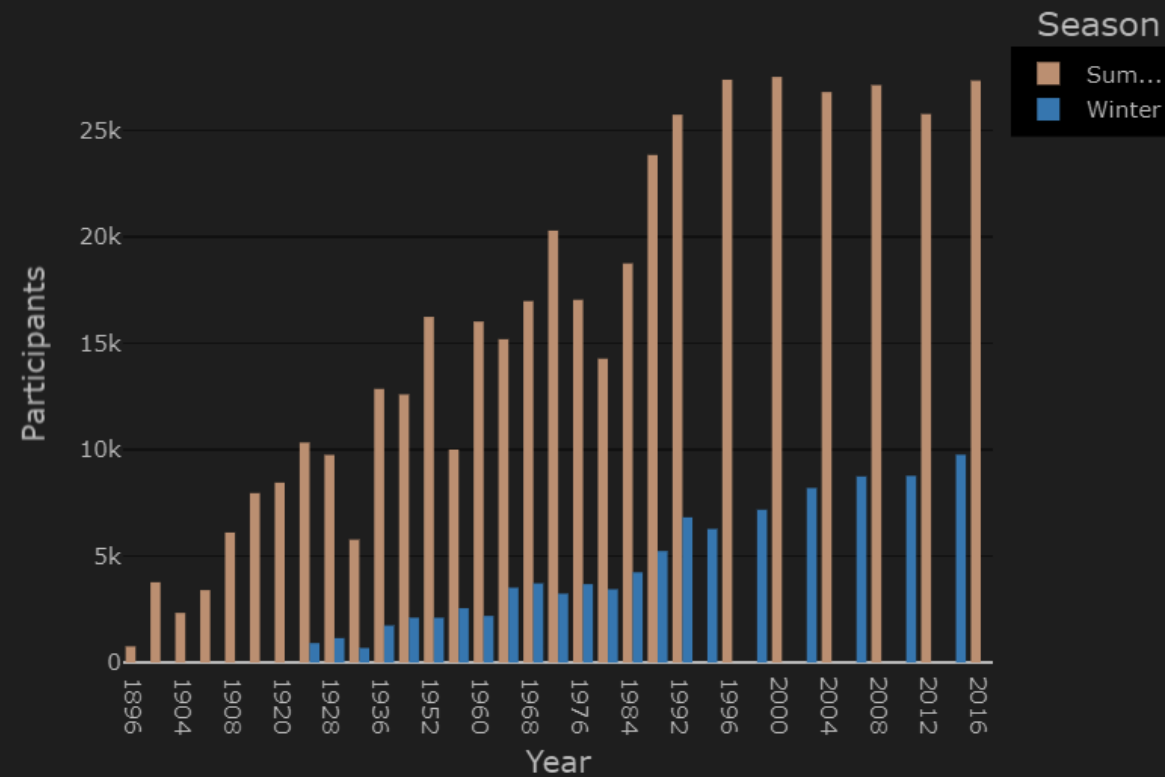
## ➤ APPROACH


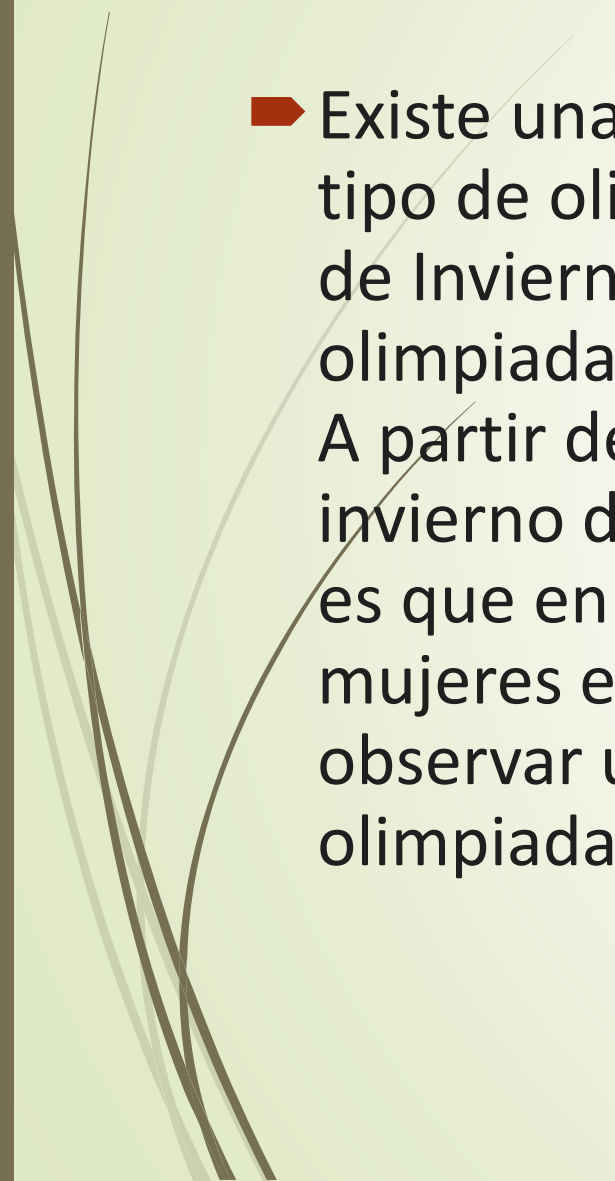
- Se harán uso de los campos sex, age, NOC, games, year, event, medal para la mayoría de las hipótesis destacadas

# Correlations

```
SELECT Season, Year, Count(*) as Participants
FROM athlete_events
WHERE Year IS NOT NULL
GROUP BY Season, Year
ORDER BY Year ASC
```

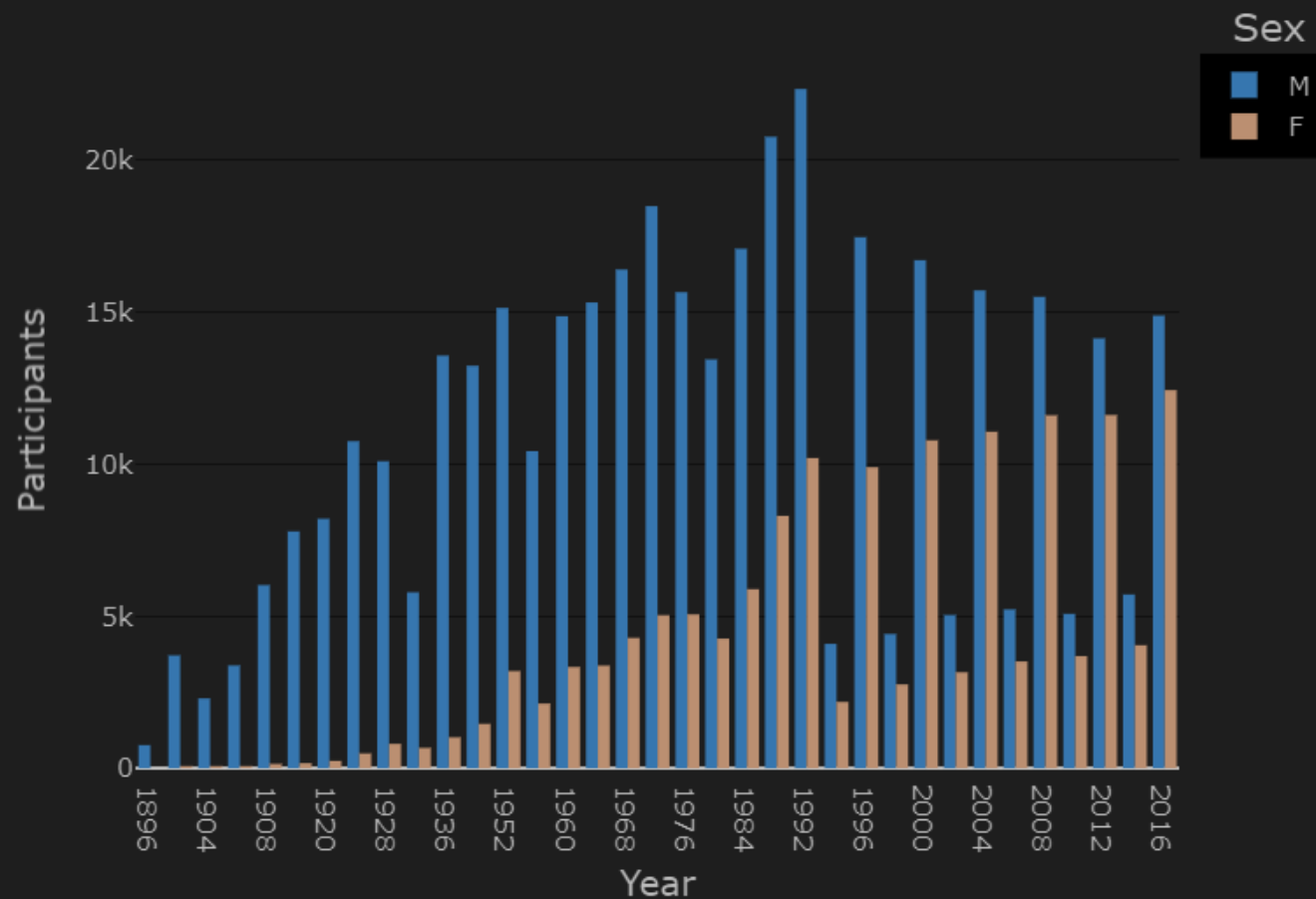
► (2) Spark Jobs


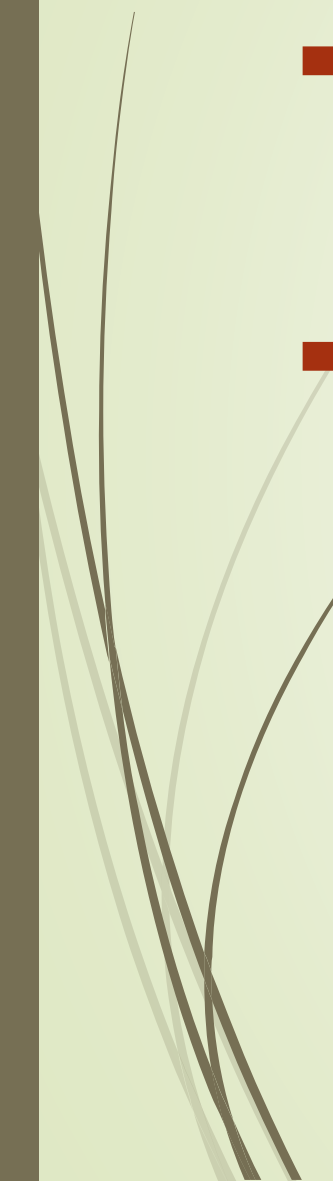


- 
- 
- Existe una correlación entre los años de las competencias y el tipo de olimpiada realizada (Olimpiadas de verano y Olimpiadas de Invierno). Desde el inicio de las olimpiadas hasta 1992, las olimpiadas de verano y de invierno se realizaban en el mismo año. A partir de las olimpiadas de 1994, las Olimpiadas de verano y de invierno decidieron hacerse de forma alternada. Otra correlación es que en los primeros juegos olímpicos, la participación de las mujeres era mínima o nula. A partir del año 1960 se puede observar un incremento en la participación femenina en las olimpiadas.

```
SELECT Sex, Year, Count(*) as Participants
FROM athlete_events
WHERE Year IS NOT NULL
AND Sex IN ('M','F')
GROUP BY Sex, Year
ORDER BY Year ASC
```

► (2) Spark Jobs



- 
- 
- ➡ -> Se observan que muchas de las olimpiadas realizadas en verano cuentan con una gran participación de mujeres comparadas con las olimpiadas de invierno a partir de 1994 en adelante.
  - ➡ -> Se realizó un análisis utilizando gráficas para poder realizar regresiones lineales entre los países participantes y las cantidades de medallas recibidas en las olimpiadas por año para poder predecir los países que sí o sí se llevan algunas medallas a casa.

# NEW METRICS

- Se creo una metrica llamada "Participantes/Season", para rastrear la cantidad de participantes por olimpiada realizada en cualquier temporada del año (Invierno o Verano)

► (2) Spark Jobs

	Season ▲	Year ▲	Participants ▲
1	Summer	1896	760
2	Summer	1900	3774
3	Summer	1904	2332
4	Summer	1906	3404
5	Summer	1908	6108
6	Summer	1912	7960
7	Summer	1920	8450

Showing all 51 rows.

Command took 4.58 seconds -- by juanvillapre2230@gmail.com at 15/11/2021 17:46