Análisis del conjunto de datos de los atletas olímpicos

El conjunto de datos utilizado para este proyecto se encuentra en el siguiente link.

Descripción del conjunto

La información del conjunto de datos corresponde desde el año 1896 hasta el año 2016. Cada una de las filas corresponde a la información de un atleta en específico.

Los objetivos para este análisis son los siguientes:

- Examinar y limpiar el conjunto de datos
- Explorar distribuciones numéricas y categóricas utilizando estadística y gráficos
- Explorar relaciones de múltiples características mediante estadísticas y gráficos

El conjunto de datos cuenta con *271,116* filas y *14* columnas de datos los cuales son datos de los distintos juego olímpicos jugados desde el año *1896* hasta *2016*. Al realizar un análisis general se descubrio que varias columnas del conjunto tenian datos faltantes, los cuales podrían afectar a los resultados finales que se podrían llegar a obtener.

Las columnas que presentan datos nulos son las siguientes:

- Age
- Height
- Weight
- Medal

Tratamiento de los datos nulos

Como primer tarea será tratar los datos nulos del conjunto llenándolos de información que no afecte al resultado final obtenido, para rellenar los datos nulos de las columnas numéricas se tomó la decisión de llenarlos del promedio de esa columna por la siguiente razón.

Si tenemos la siguiente *Serie* de datos:

```
[1, 2, NaN, 4, 5]
```

Al momento de calcular el promedio de esa *Serie* con *Python* obtenemos el siguiente resultado: `3.0`. Este resultado se debe a que, en *Python*, al momento de calcular el promedio de una *Serie* que cuenta con un dato nulo éste no lo toma en cuenta y solamente hace la

operación con los demás datos, en este caso `1, 2, 4, 5`. Al sumarlos y dividir entre la longitud de la *Serie* obtenemos el resultado `3.0`.

Es importante hacer el tratamiento de esos datos faltantes en las colecciones de datos que estemos trabajando ya que, si no lo hacemos, al momento de continuar con nuestro análisis podríamos perder información de las filas en donde se encuentran esos datos nulos. Por esta razón, vamos a buscar la mejor forma de llenar esos datos faltantes con valores que se puedan utilizar.

Lo primero que debemos buscar, es llenar ese dato faltante con un valor que no altere los resultados de las distintas operaciones que se puedan llegar a hacer sobre esa colección de datos, por ejemplo, que pasaría si nosotros cambiamos los valores nulos con ceros (0):

```
[1, 2, 0, 4, 5]
```

Si nosotros hacemos eso y volvemos a calcular el promedio, veremos que ahora tenemos el siguiente resultado, `2.4`.

Este resultado está mal, ya que ahora, *Python* toma toda la colección de datos, incluyendo al cero, y lo divide entre la longitud de la colección, la cual pasó de `4` a `5`. Agregar el cero no cambia el resultado de la suma, pero sí cambia la longitud de la *Serie* con la que se va a hacer la división del promedio, por lo que usar el cero como cambio en los datos nulos no es la mejor manera.

Otra mejor forma de cambiar esos datos nulos por un valor que se pueda usar y que no altere a la colección como si lo hace el cero, es llenarlos por el promedio de la colección original, por ejemplo, sabemos que el promedio de la siguiente colección es `3.0`.

Ahora, si usamos ese promedio como sustituto de los valores nulos en la *Serie* de la siguiente forma:

```
[1, 2, 3, 4, 5]
```

El promedio de esa nueva colección seguirá siendo `3.0`, con esto no alteramos el resultado que podemos obtener de esa colección.

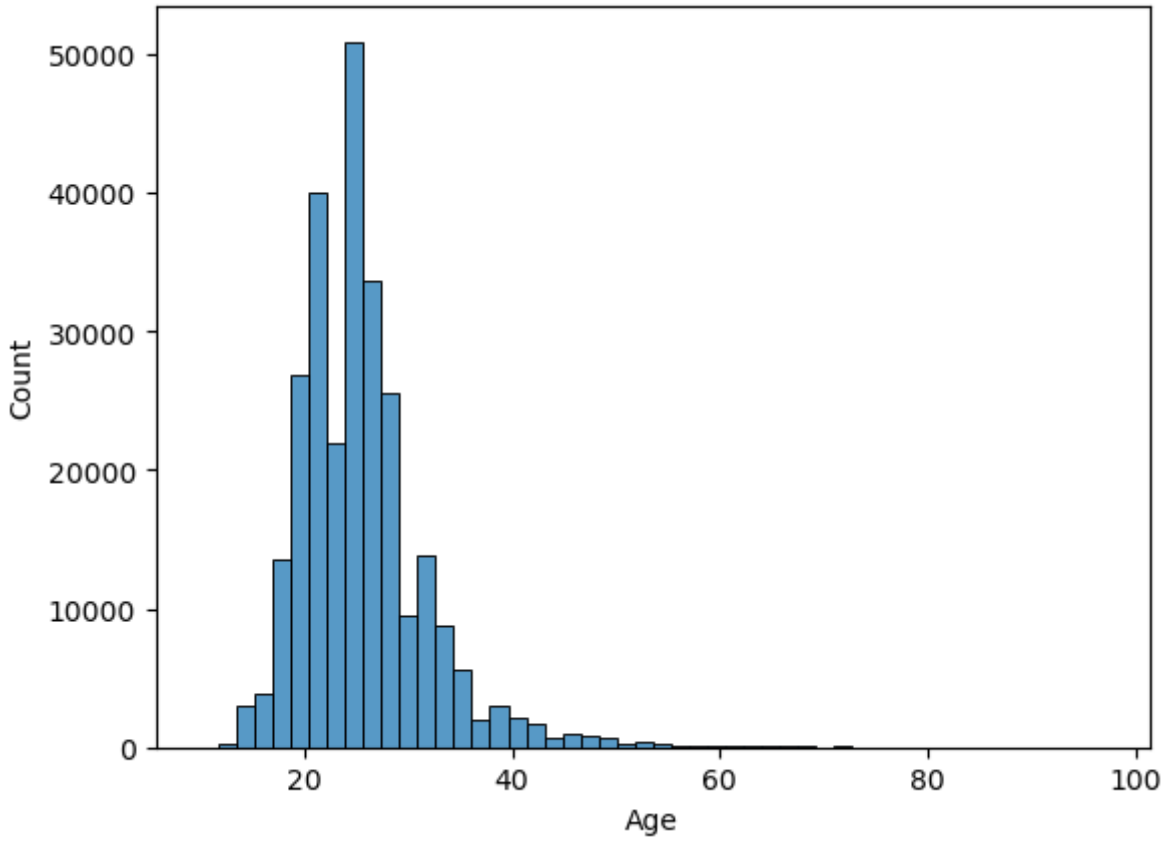
Para la columna `Medal` que almacena texto, simplemente cambiaremos los valores nulos por el texto `NA`

De ésta forma conseguimos que todas las columnas tengan la misma cantidad de datos, que son *271,116*.

Análisis Exploratorio de los Datos

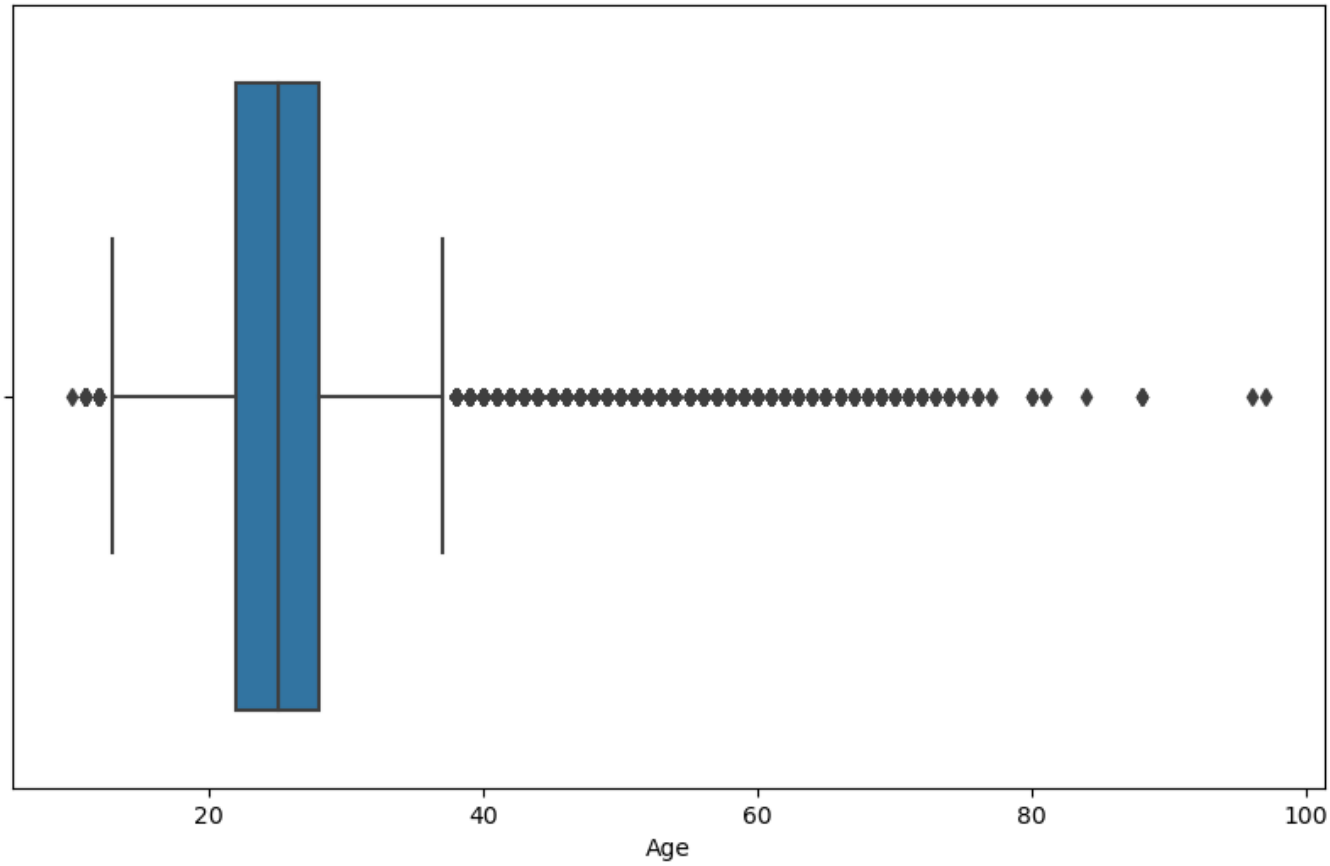
Lo siguiente que haremos sobre nuestro conjunto de datos será obtener una serie de gráficas que nos ayudarán a obtener un vistazo general del conjunto para saber que camino tomar al momento de desarrollar nuestro análisis.

Como primer punto dibujaremos un *histograma* con el que veremos la distribución de la edad de los atletas.



Podemos ver que la gran mayoría de los datos se encuentran entre los 20 y 40 años, aunque podemos ver que el histograma dibuja en el eje de las *x* hasta el 100, lo que quiere decir que enste conjunto de datos se encuentran algunos atletas que participaron en los juegos olímpicos llegando hasta esa edad.

Para comprender mejor la distribución de los datos, dibujaremos un *boxplot* para ver como están distribuidos los datos.



En el *boxplot* nos da distintos resúmenes estadísticos que nos sirven para entender la distribución de los datos, entre éstos podemos encontrar los siguientes:

- En la caja se encuentran la mayoría de los datos, es esta caja podemos encontrar el 50% de los datos de nuestro conjunto. Podemos comprobar esto si vemos el histograma.
- Los "bigotes" de la caja representan el los datos que no se consideran atípicos, en este caso los bigotes tienen un valor de 1.5 * IQR (El rango intercuartílico). En este caso los bigotes tienen los siguientes valores:
  - 28 + (1.5 * 6) = 37
  - 22 - (1.5 * 6) = 13

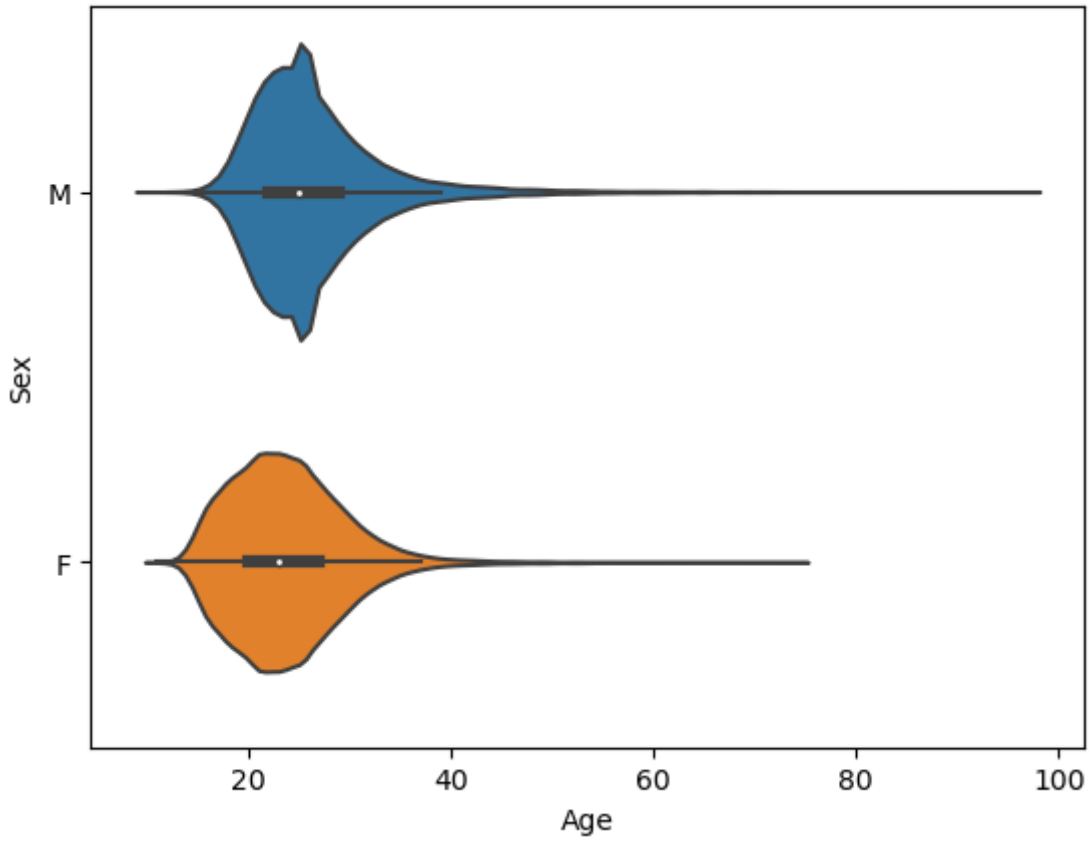
  Cualquier dato que éste por encima de *37* o por debajo de *13* se considerará atípico.

Con este gráfico podemos ver que existen muchos deportistas muy viejos que han participado en los juegos olímpicos, por lo que ahora buscaremos más información sobre aquellos deportes en los que participaron éstos deportistas.

Ahora que conocemos los valores en donde se consdieran atípicas las edades, haremos un filtrado por esas edades para ver información más específica de cada uno de ellos. Con este filtrado encontramos lo siguiente:

- La cantidad de datos de deportistas mayores a 37 años son *11,928*
- La cantidad de datos de deportistas menores a 13 años son *53*

Podemos ver otro gráfico que nos ayuda a comprender mejor la distribución de los datos que es muy parecida al *boxplot* dibujado anteriormente, este gráfico de violín combina los elementos de un gráfico *boxplot* con una visualización de la distribución de los datos.



En el anterior gráfico vemos lo siguiente:

- La forma del violín muestra la distribución de los datos. Es más ancho donde hay más puntos de datos y más estrecho donde hay menos.
- En el centro de cada violín, podemos ver una línea blanca la cual representa la mediana de los datos.
- Si el violín está asimétrico, puede indicar una distribución sesgada. Si es simétrico podría indicar una distribución más normal.

Sabiendo esto, buscaremos los principales deportes que practican estos atletas:

Atletas menores

Deporte	Cantidad
Natación	25
Patinaje Artístico	15
Remo	5
Gimnasia	5

Deporte	Cantidad
Atletismo	2
Buceo	1

Atletas mayores

Deporte	Cantidad
Tiro	3178
Arte	2226
Equitación	1997
Navegación	1040
Esgrima	1031

Podemos notar aquí una diferencia, y es que los deportes en donde se tienen a atletas menores son deportes en los que la actividad física es mucho mayor.

Las siguientes tablas los principales datos en las distintas variables, estas tablas nos van a servir para conocer al conjunto de forma general.

Información en las distintas categorías

Ahora buscaremos información más específica del conjunto de datos, como primer paso será conocer la distribución en distintas variables dentro del conjunto.

Hombres y Mujeres

Sexo	Cantidad
Hombre	196,594
Mujer	74,522

Atletas en las distintas temporadas

Temporada	Cantidad
Summer	222, 552
Winter	48, 564

Países con más atletas

País	Cantidad
Estados Unidos	17,847
Francia	11,988
Gran Bretaña	11,404
Italia	10,260
Alemania	9,326

Año y temporada que más atletas tuvo

Temporada	Cantidad
2000 Summer	13821
1996 Summer	13780
2016 Summer	13688
2008 Summer	13602
2004 Summer	13443

Ciudad que más atletas ha recibido

Ciudad	Cantidad
Londres	22,426
Atenas	15,556
Sídney	13,821
Atlanta	13,780
Rio de Janeiro	13,688

Deportes con más atletas

Deporte	Cantidad
Atletismo	38624
Gimnasia	26707
Natación	23195
Tiro	11448
Ciclismo	10859

Eventos que más atletas han tenido

Evento	Cantidad
Futbol masculino	5,733
Hockey sobre hielo masculino	4,762
Hockey masculino	3,958
Polo acuático masculino	3,358
Basquebol masculino	3,280

Cantidad de medallas

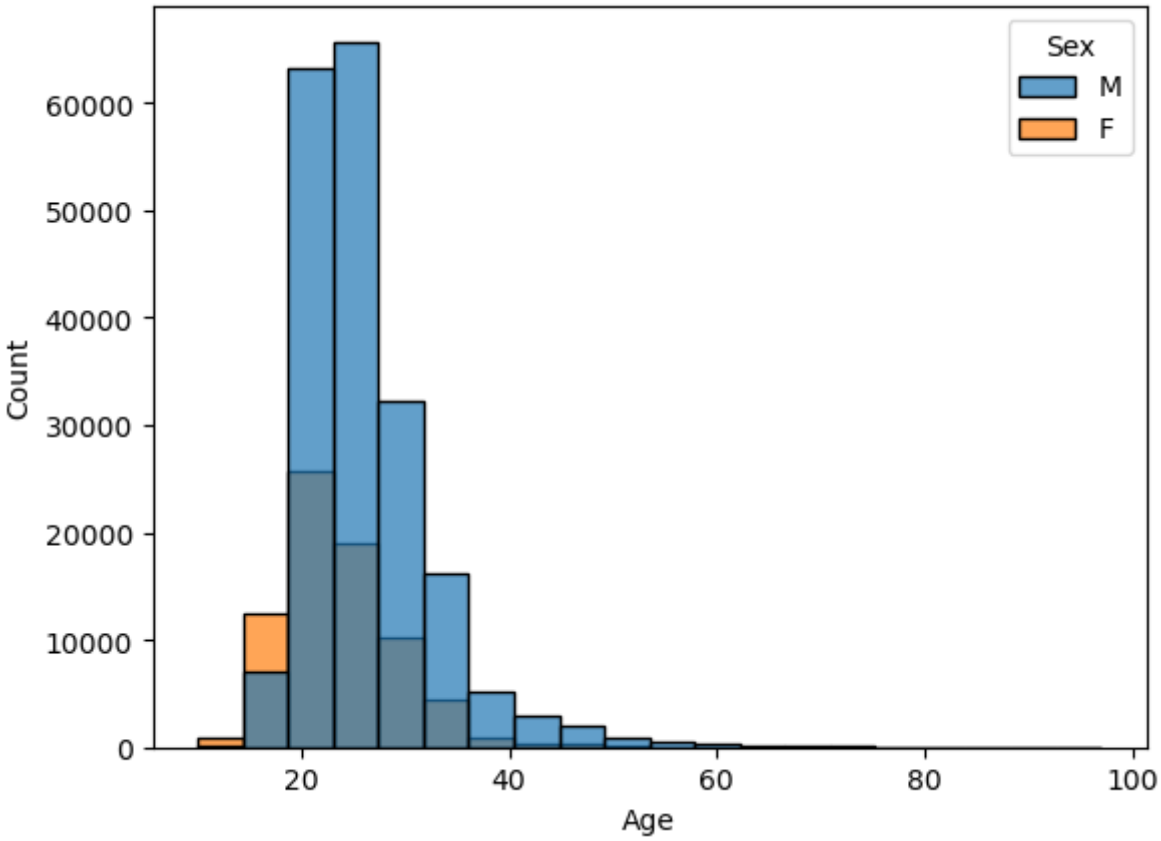
Medallas	Cantidad
Sin dato	231,333
Oro	13,372
Bronce	13,295
Plata	13,116

Información de los atletas

En la siguiente tabla se encuentra el promedio de la edad, peso y estatura de todos los atletas a lo largo de los juegos olímpicos.

Sexo	Edad	Estatura	Peso
F	24	168	61
M	26	178	74

Podemos ver que en general los hombres presentes en los juegos olímpicos son más altos y tienen un mayor peso. En el siguiente histograma podemos ver la distribución de las edades de los atletas entre hombre y mujeres para mostrar la disparidad entre ambos.



Con esto podemos comprobar que no hay igualdad de condiciones, al menos históricamente en las oportunidades que han tenido los hombres y las mujeres de pode competir en los Juegos Olímpicos. Para conocer mejor la distribución de hombres y mujeres en esta competición, en la siguiente gráfica se muestran la cantidad de equipos, deportes y eventos que han tenido los hombres y las mujeres en las distintas temporadas.

Season	Sex	Team	Sport	Event
Summer	F	352	40	214
	M	1118	49	491
Winter	F	144	14	57
	M	214	17	67

Podemos ver en la pasada tabla que los Juegos Olímpicos de verano existe la mayor diferencia entre la cantidad de equipos masculinos compitiendo así como en los eventos. Esta diferencia es menos notoria en los Juegos Olímpicos de inverno pero, aunque estos Juegos Olímpicos lleven existiendo menos tiempo, es notoria la diferencia entre las mujeres y hombres que participan.

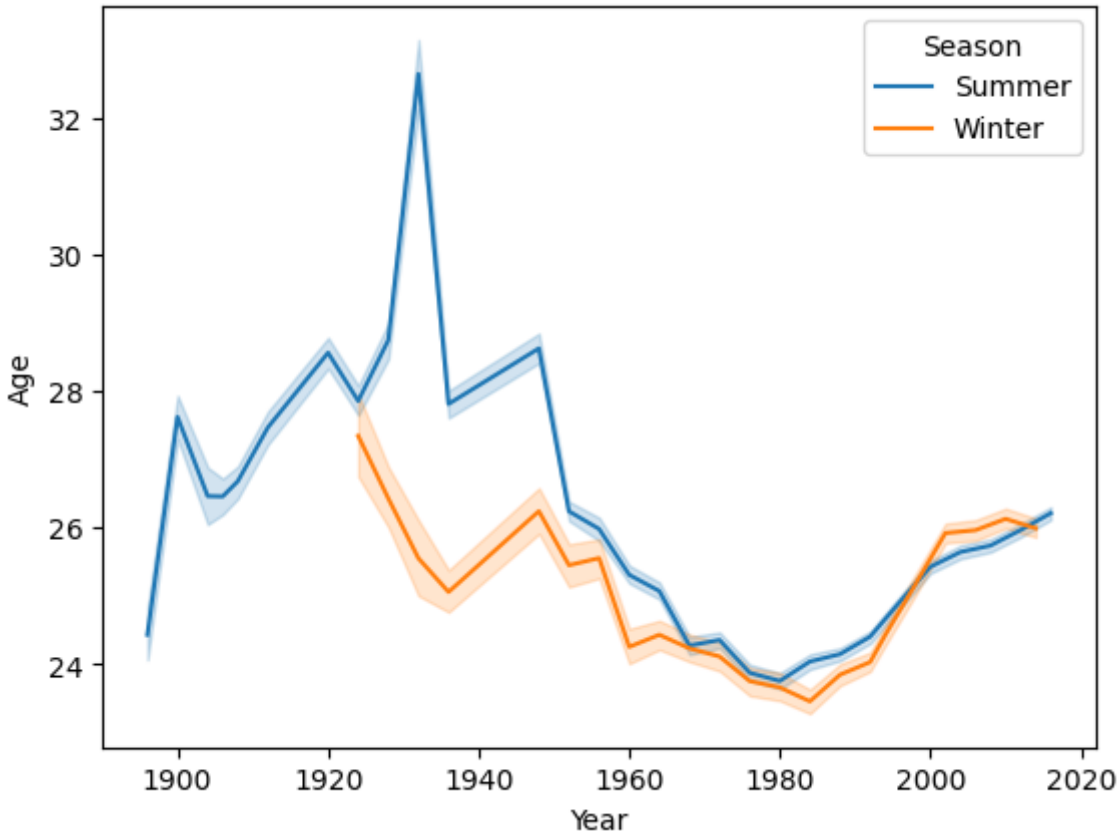
Finalmente en la información de los atletas que ganan medallas podemos encontrar que el promedio general de la edad, estatura y peso de todos ellos es muy parecida, encontrando que

en general los atletas que ganan medallas son muy jóvenes. En la siguiente tabla podemos ver esa información desglosada:

Medal	Season	Sex	Age	Height	Weight
Bronze	Summer	F	24.63	171.18	64.05
		M	26.32	179.43	76.39
	Winter	F	25.12	167.40	61.38
		M	26.38	178.89	77.10
Gold	Summer	F	24.21	171.67	64.38
		M	26.47	179.87	76.87
	Winter	F	25.20	167.62	62.43
		M	26.60	179.54	77.78
NA	Summer	F	23.54	168.36	60.59
		M	26.38	177.59	73.74
	Winter	F	23.86	167.41	60.89
		M	25.38	177.84	74.83
Silver	Summer	F	24.29	171.39	64.06
		M	26.63	179.48	76.49
	Winter	F	25.24	167.97	62.26
		M	26.43	179.09	77.25

- El rango de edades de las personas que que ganan medallas se encuentra entre 24 a 26 años generalmente.
- La estatura de esos atletas se encuentra entre 167 a 179 cm
- El peso de los atletas ganadores de alguna medalla no supera los 77 Kg

Sabiendo que los atletas ganadores son muy jóvenes, dibujaremos una gráfica para ver la edad de los atletas a lo largo de los años en las distintas temporadas.



Vemos que a lo largo de los años la edad de los atletas siempre ha estado por debajo de los 30 años. Aunque después del 1940 la edad promedio de los atletas ha ido disminuyendo, el punto mínimo se alcanzo en 1980 y, a partir de ese punto, la edad promedio de los atletas ha vuelto a subir.

Conclusiones

Históricamente los Juegos Olímpicos han sido unos deportes en donde los hombres han tenido principal protagonismo en las distintas competiciones el cual no ha cambiado mucho con el paso de los años ya que, si revisamos la cantidad de atletas masculinos contra las atletas femeninas vemos una diferencia de 122,075 atletas, o lo que es igual una diferencia del 62.09%. Algo importante a tener en cuenta.
Pasando ahora a la información en base a la edad notamos que los atletas que pasan la edad común que se tiene en las distintas competencias (siendo la edad máxima los 37 años) la mayoría de estos atletas compiten en pruebas que no necesitan un esfuerzo físico como lo es Tiro, Arte o Equitación. A diferencia de los atletas que presentan una edad menor a lo normal (siendo la edad mínima los 13 años), estos si que participan en pruebas que tienen una carga física mayor como Natación, Patinaje Artístico o Remo. Teniendo en cuenta esta información es más común encontrarse con atletas viejos que atletas jóvenes.
Observamos que la mayoría de los atletas que participan en los Juegos Olímpicos tienen edad que no superan los 28 años y podemos ver una relación de la edad con distintos datos obtenidos del conjunto, los cuales se muestran a continuación:

- Los deportes que presentan la mayoría de atletas son aquellos en donde el esfuerzo físico necesario es muy alto, a continuación se muestran los cinco deportes más populares:
  - Atletismo
  - Gimnasia
  - Natación
  - tiro
  - Ciclismo
- La edad promedio de los atletas ganadores de alguna medalla no superan los 26 años, lo que indica que en general, los atletas en estos momentos se encuentran en su máximo estado físico que les permite ganar alguna medalla en las distintas competiciones dentro de los juegos olímpicos.
- Podemos ver como la edad máxima de los competidores cambia cuando se habla de las distintas temporadas, en la temporada de verano vemos que la edad máxima de un atleta presente fue de 97 años, a diferencia de cuando se hablan de los juegos de invierno en donde la edad máxima fue de 58 años., ambos masculinos.
  Este comportamiento de la edad de los atletas podemos ver históricamente como es que se llego a este punto, ya que si vemos la gráfica de la edad de los atletas a lo largo del tiempo. En sus comienzos los Juegos Olímpicos comenzaron con atletas jóvenes pero rápidamente la edad fue subiendo alcanzando su máximo aproximadamente en 1930. Después de este año la edad comenzó a bajar teniendo su pico más bajo en 1980 para volver a subir llegando ahora a un promedio de 26 años en la actualidad la edad de los distintos atletas.
  Podemos buscar distintos motivos del porque la edad bajó tanto después de 1940, una de las razones por las que la edad de los atletas haya bajo tan drásticamente es a la Segunda Guerra Mundial que acabo con la vida de muchas