# JUEGOS OLÍMPICOS

**PREGUNTAS:** ¿Cómo ha sido el desempeño de Colombia en los juegos Olímpicos de verano desde su primera participación? ¿Qué deporte se debería incentivar más en el país para poder conseguir mejores resultados en los juegos olímpicos? ¿Cómo ha sido el rendimiento de la delegación de weightlifting(levantamiento de pesas) colombiana a comparación de Rusia? ¿Dado N(cantidad determinada) atletas que compitan en representación de un NOC, cuantos de ellos podrían obtener una medalla de cualquier tipo?

## 1. COMPRENSIÓN DEL NEGOCIO

- **Determinar objetivo de negocio:** Conocer el trayecto histórico de Colombia frente a los juegos Olímpicos y cuales son sus principales fortalezas.
- **Evaluación de la situación:** A partir de un dataset que contiene 271mil observaciones sobre participantes en los juegos Olímpicos de los últimos 120 años y 15 variables que describen la información de los atletas como ID ( Número único de c/u), Nombre, género, Edad, Altura, Peso,Nombre del equipo, NOC (código de tres letras del comité Nacional Olímpico, Juego,Año, Temporada, Ciudad, Deporte, Evento y Medallas. Los datos fueron extraídos y ordenados a partir de www.sports-reference.com.
- **Determinar objetivos de la minería de datos:** Comparar las variables de desempeño individual a través de los años, eventos y resultados en relación con Colombia y el mundo.
- **Producir un plan de trabajo:**
  Recolección, clasificación y compresión de datos
  Ordenar las observaciones según el desempeño en cantidad de medallas
  Comparar Rendimiento entre países
  Analizar la diferencia entre medallistas y no medallistas
  Predecir cantidad de medallistas

## 2. COMPRENSIÓN DE LOS DATOS

- **Recolectar los datos iniciales:** Los datos fueron tomados a partir de un dataset de la pagina web Kaggle (https://www.kaggle.com/) el cual lleva el titulo de “120 años de historia olímpica: atletas y resultados”, publicado por el científico de datos Randi H Griffin que a su vez obtuvo la información sin formato u orden de la página Sports Reference (www.sports-reference.com ).
- **Describir los datos:** El documento tipo csv contiene 271116 filas y 15 columnas, las columnas describen las características individuales de un participante en un evento en específico y las columnas se encuentran organizadas ascendentemente por ID, clave en la organización y visualización de la totalidad del Data Frame.
  Las variables que se encuentran son.
  -ID el cual es un número entero único por atleta que ayuda a identificarlo entre los eventos y/o años.
  -Name, en forma de string que almacena Nombre y Apellido del atleta
  -Sex, identificado por la letra M (Male) o F (Female)

-Agedel participante
-Height, altura en centímetros
-Weight, peso en kilogramos
-Team, el nombre del equipo/país
-NOC, es la nomenclatura oficial otorgada por el comité olímpico a cada país en un código de tres letras.
-Games, año y temporada del evento
-Year, año en forma de número entero
-Season, temporada en la que sucedió el evento, dividido entre juegos de verano (Summer) y juegos de invierno (Winter)
-City, Ciudad de nacimiento del atleta
-Sport, deporte en el que se desempeña el atleta
-Event, Eventos por deporte
-Medal, Medalla otorgada al participante, se divide en Oro, plata, bronce y "Sin medalla"

- **Explorar los datos:**

shape + info() original

```
(206165, 15)

<class 'pandas.core.frame.DataFrame'>
Int64Index: 206165 entries, 0 to 271115
Data columns (total 15 columns):
ID        206165 non-null int64
Name      206165 non-null object
Sex       206165 non-null object
Age       206165 non-null float64
Height    206165 non-null float64
Weight    206165 non-null float64
Team      206165 non-null object
NOC       206165 non-null object
Games     206165 non-null object
Year      206165 non-null int64
Season    206165 non-null object
City      206165 non-null object
Sport     206165 non-null object
Event     206165 non-null object
Medal     30181 non-null object
dtypes: float64(3), int64(2), object(10)
memory usage: 25.2+ MB
```

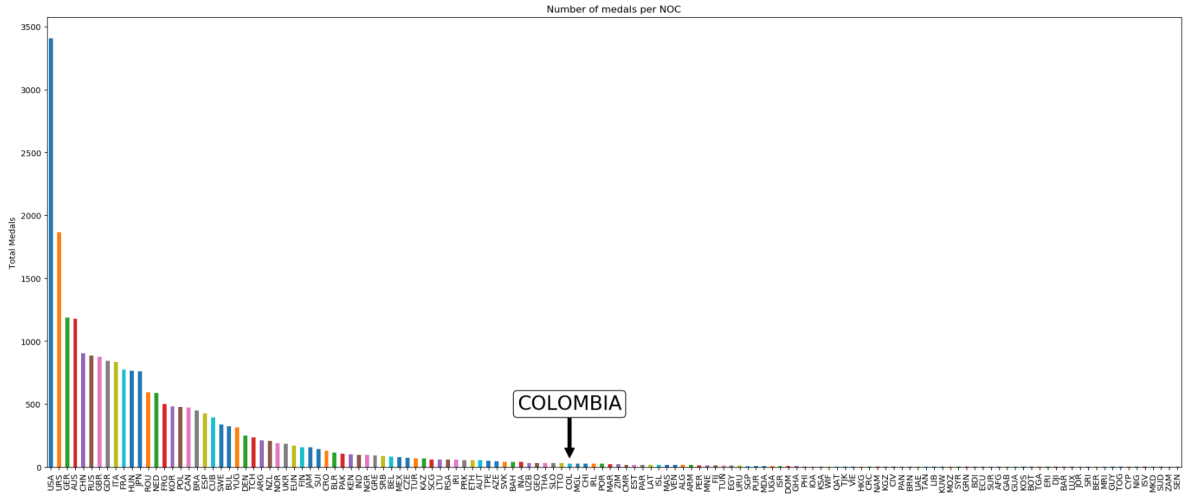
head() original

| | ID | Name | Sex | Age | Height | Weight | Team | NOC | Games | Year | Season | City | Sport | Event | Medal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | A Dijiang | M | 24.0 | 180.0 | 80.0 | China | CHN | 1992 Summer | 1992 | Summer | Barcelona | Basketball | Basketball Men's Basketball | NaN |
| **1** | 2 | A Lamusi | M | 23.0 | 170.0 | 60.0 | China | CHN | 2012 Summer | 2012 | Summer | London | Judo | Judo Men's Extra-Lightweight | NaN |
| **4** | 5 | Christine Jacoba Aaftink | F | 21.0 | 185.0 | 82.0 | Netherlands | NED | 1988 Winter | 1988 | Winter | Calgary | Speed Skating | Speed Skating Women's 500 metres | NaN |
| **5** | 5 | Christine Jacoba Aaftink | F | 21.0 | 185.0 | 82.0 | Netherlands | NED | 1988 Winter | 1988 | Winter | Calgary | Speed Skating | Speed Skating Women's 1,000 metres | NaN |
| **6** | 5 | Christine Jacoba Aaftink | F | 25.0 | 185.0 | 82.0 | Netherlands | NED | 1992 Winter | 1992 | Winter | Albertville | Speed Skating | Speed Skating Women's 500 metres | NaN |

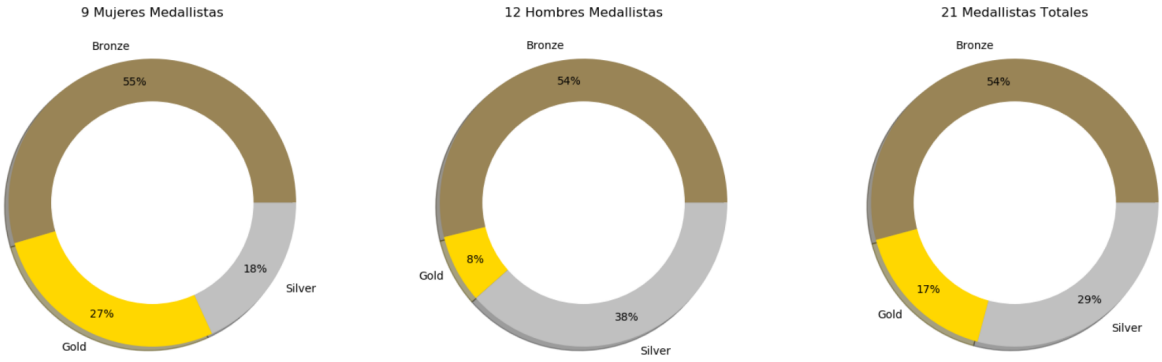
describe() original

| | ID | Age | Height | Weight | Year |
|---|---|---|---|---|---|
| **count** | 206165.000000 | 206165.000000 | 206165.000000 | 206165.000000 | 206165.000000 |
| **mean** | 68616.017675 | 25.055509 | 175.371950 | 70.688337 | 1989.674678 |
| **std** | 38996.514355 | 5.483096 | 10.546088 | 14.340338 | 20.130865 |
| **min** | 1.000000 | 11.000000 | 127.000000 | 25.000000 | 1896.000000 |
| **25%** | 35194.000000 | 21.000000 | 168.000000 | 60.000000 | 1976.000000 |
| **50%** | 68629.000000 | 24.000000 | 175.000000 | 70.000000 | 1992.000000 |
| **75%** | 102313.000000 | 28.000000 | 183.000000 | 79.000000 | 2006.000000 |
| **max** | 135571.000000 | 71.000000 | 226.000000 | 214.000000 | 2016.000000 |

Cantidad de medallas en Colombia vs resto del mundo




Cantidad de medallistas en Colombia por sexo



- **Verificar la calidad de los datos:** Se corroboró la veracidad de los datos al comparar la fuente principal (www.sports-reference.com) con documentación web sobre la cantidad de medallas, países participantes y datos atípicos que parecen errores y no datos verídicos, por ejemplo, se encuentran participantes jóvenes con una edad mínima de 11 años o de avanzada edad hasta los 97 años, ambos son casos reales de atletas en gimnasia y equitación respectivamente.

**3. PREPARAR LOS DATOS**

- **Seleccionar:** En cuanto a variables todas se conservan, ya que cada una de estas es pertinente para diferentes análisis del dataset. “Year”, “NOC”, “ID”, “Sport”,”Medals” son las columnas que más se destacan ya que permiten hacer análisis del rendimiento de los atletas a través de los años, según las medallas, el país o el deporte; las demás variables permiten complementar el tipo de atleta que participan en los eventos olímpicos.
- **Limpiar:**Inicialmente las observaciones se organizaron por año de forma ascendente lo que da una perspectiva de cambio a través del tiempo y también se realizó una limpieza de una minoría de atletas que les faltaba información de peso, altura o deporte, por lo cual se sacaron del dataset, ya que generaban más incongruencias que información general.
- **Construir:** Los valores predictores son cantidad de asistentes a los juegos olímpicos contra aquellos que ganaron medalla, lo que permite visualizar una tasa de éxito en los juegos según la cantidad de asistentes en un deporte. Para esto se sacaron dos dataframes diferentes uno agrupando los “ID” únicos por año y el otro “Medallistas” agrupados por año, luego al obtener la cantidad de los dos “por año” se eligen las dos variables para generar una predicción.
- **Integrar:** Para responder las preguntas descriptivas era necesario centrar el data frame en Colombia y posteriormente en el deporte “Weightlifting” lo que permite graficar el rendimiento de Colombia en los Olímpicos.
- **Dar formato:** La variable de “Medals” le asignaba a los no ganadores de medallas un valor Nulo, para términos de mejor visualización se cambiaron todos aquellos nulos por “No medal”; también se renombraron algunas columnas para especificar las unidades de medida que manejan los datos, estas fueron “Age”, “Height”, “Weight” y se cambiaron por “Age (Years)”, “Height (cm)”, “Weight (kg)” . Todos los demás valores string del dataset tienen el formato adecuado.

**4. MODELAR**

- **Seleccionar los modelos:** El modelo de ciencia de datos escogido es Regresión lineal ya que el problema de negocio es de carácter predictivo. La regresión lineal es una técnica de análisis predictivo básica que utiliza datos históricos para predecir una variable de salida.

- **Diseñar pruebas:** Hay dos tipos de variables en un modelo de regresión lineal:

  -La variable de entrada o predictor es la variable (s) que ayuda a predecir el valor de -la variable de salida. Se conoce comúnmente como X .
  La variable de salida es la variable que queremos predecir. Se conoce comúnmente como Y .
  Para estimar Y usando regresión lineal, asumimos la ecuación:

  $Y_e = \alpha + \beta X$
  donde $Y_e$ es el valor estimado o predicho de Y basado en nuestra ecuación lineal.
  Nuestro objetivo es encontrar valores estadísticamente significativos de los parámetros α y β que minimicen la diferencia entre Y y $Y_e$.

  Si somos capaces de determinar los valores óptimos de estos dos parámetros, entonces tendremos la línea de mejor ajuste que podemos utilizar para predecir los valores de Y , dado el valor de X .

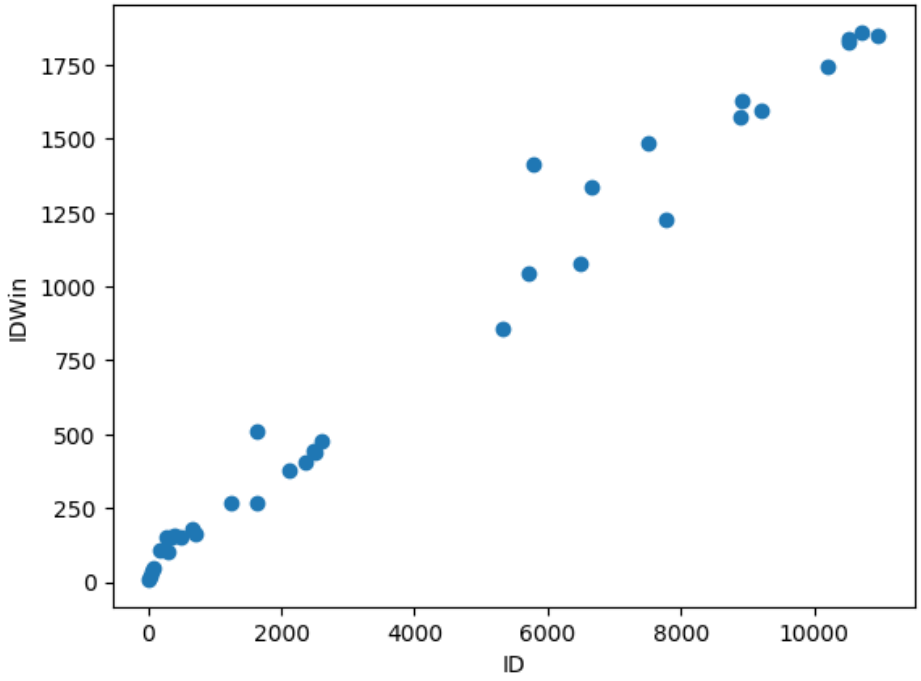
- **Desarrollar los modelos:**

  Después de realizar los dos Datasets, Cantidad de participantes por año (izquierda) y Cantidad de ganadores por año (derecha) se procede a realizar la regresión lineal entre cantidad de participantes y cantidad de medallistas.

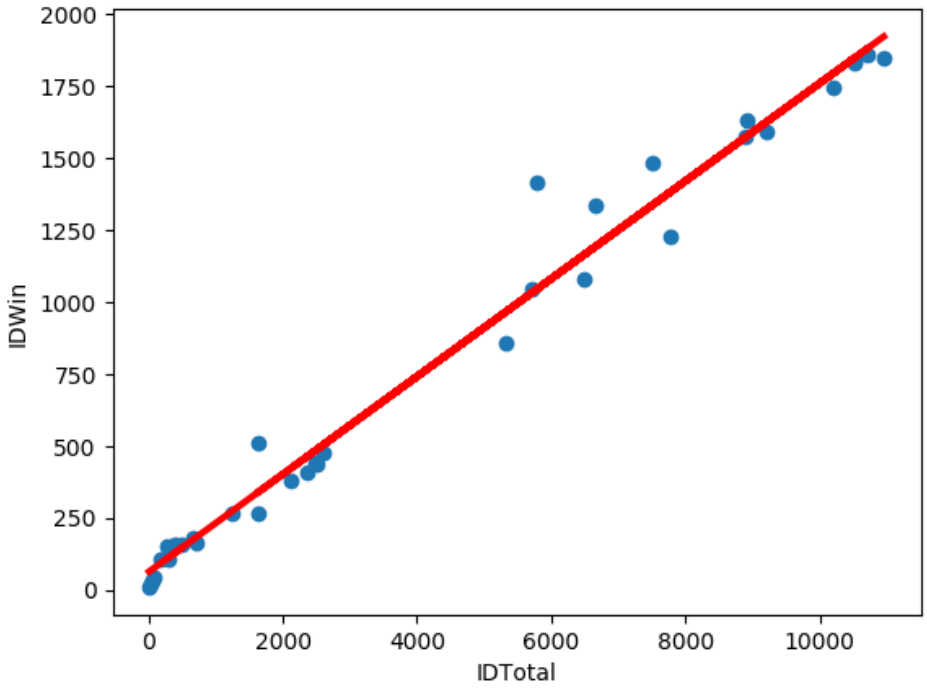
| | Year | ID |
|---|---|---|
| 0 | 1896 | 11 |
| 1 | 1900 | 34 |
| 2 | 1904 | 52 |
| 3 | 1906 | 77 |
| 4 | 1908 | 187 |
| 5 | 1912 | 302 |
| 6 | 1920 | 275 |
| 7 | 1924 | 408 |
| 8 | 1928 | 487 |
| 9 | 1932 | 348 |
| 10 | 1936 | 672 |

| | Year | ID |
|---|---|---|
| 0 | 1896 | 9 |
| 1 | 1900 | 20 |
| 2 | 1904 | 34 |
| 3 | 1906 | 45 |
| 4 | 1908 | 106 |
| 5 | 1912 | 103 |
| 6 | 1920 | 150 |
| 7 | 1924 | 158 |
| 8 | 1928 | 154 |
| 9 | 1932 | 151 |
| 10 | 1936 | 179 |

Gráfica de relación entre Cantidad de asistentes por año (Eje x) versus Cantidad de medallistas por año (Eje y)




Regresión Lineal



- **Valorar los modelos:**
  Para una cantidad de 500 participantes:

```
lr.predict([[500]])
```

  La regresión lineal predice una cantidad de medallistas de 145, con un score del 98%, es decir que el programa retorna un nivel de fiabilidad de casi el 100%

```
array([145.43792509])
```

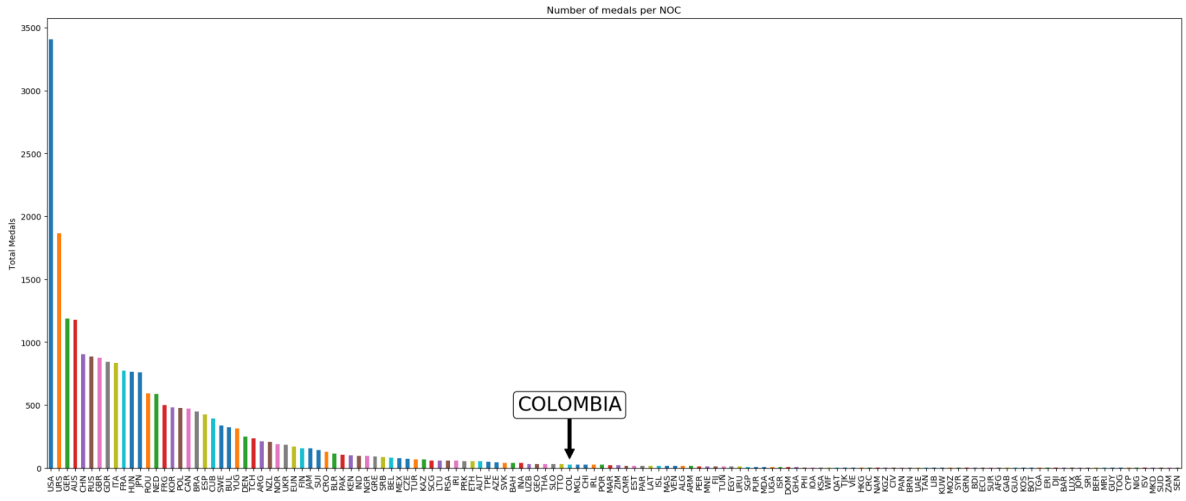
```
0.9815923529980868
```

**5. EVALUAR**

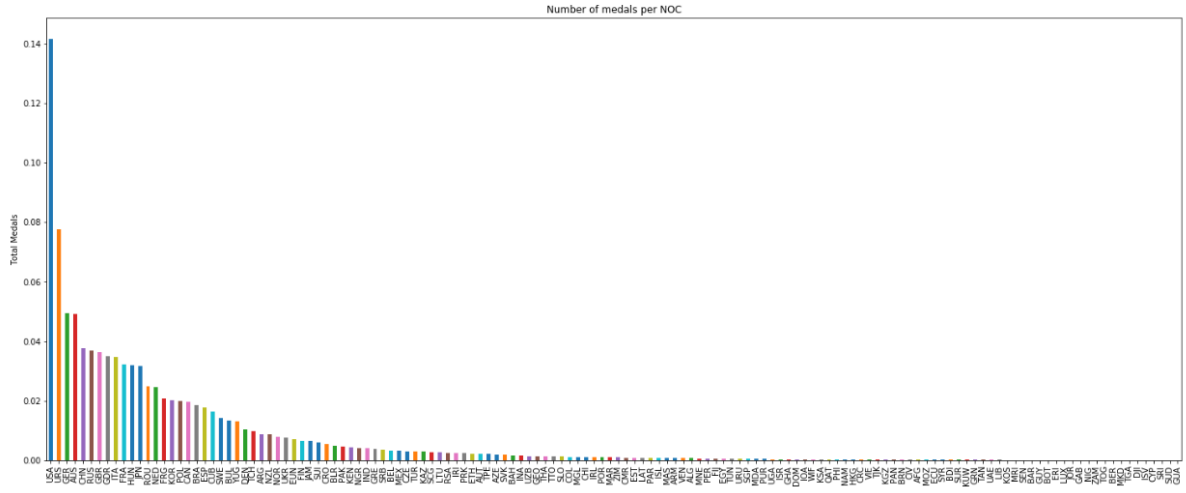
- Evaluar resultados:
  Según la pregunta ¿Dado N(cantidad determinada) atletas que compitan en representación de un NOC, cuantos de ellos podrían obtener una medalla de cualquier tipo? se obtiene un resultado satisfactorio de predicción con un score bastante alto, por lo cual se podría considerar la regresión lineal un modelo exitoso de predicción para esta pregunta de negocio.
- Revisar el proceso
  Se tuvieron en cuenta todas las variables y observaciones necesarias para realizar la predicción mediante la regresión lineal.
- Trabajo futuro
  La pregunta de negocio se planteó a nivel global dado que en Colombia hay muy poca tasa de medallistas con relación a los participantes, por lo tanto actualmente no se puede obtener una predicción nacional que potencie la posibilidad de una mayor cantidad de medallas, en el futuro con mayor cantidad de datos nacionales se logrará una predicción más pertinente a la empresa interesada.

## 6. Informe Final

### ¿Cómo ha sido el desempeño de Colombia en los juegos Olímpicos de verano desde su primera participación?




```
False    24013
True        27
Name: NOC, dtype: int64
```



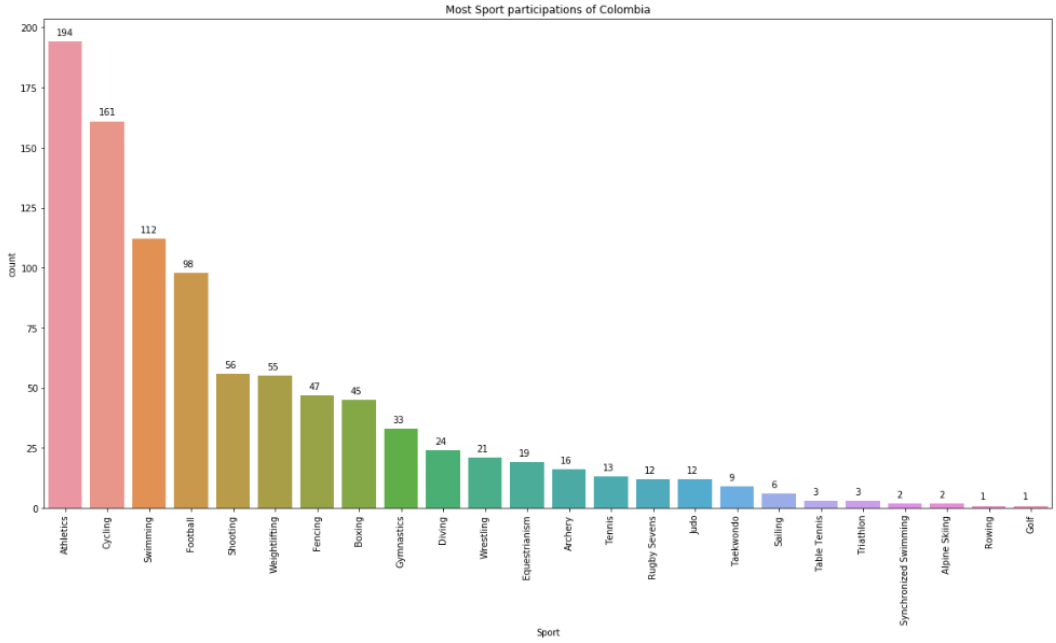

```
False    0.998877
True     0.001123
Name: NOC, dtype: float64
```

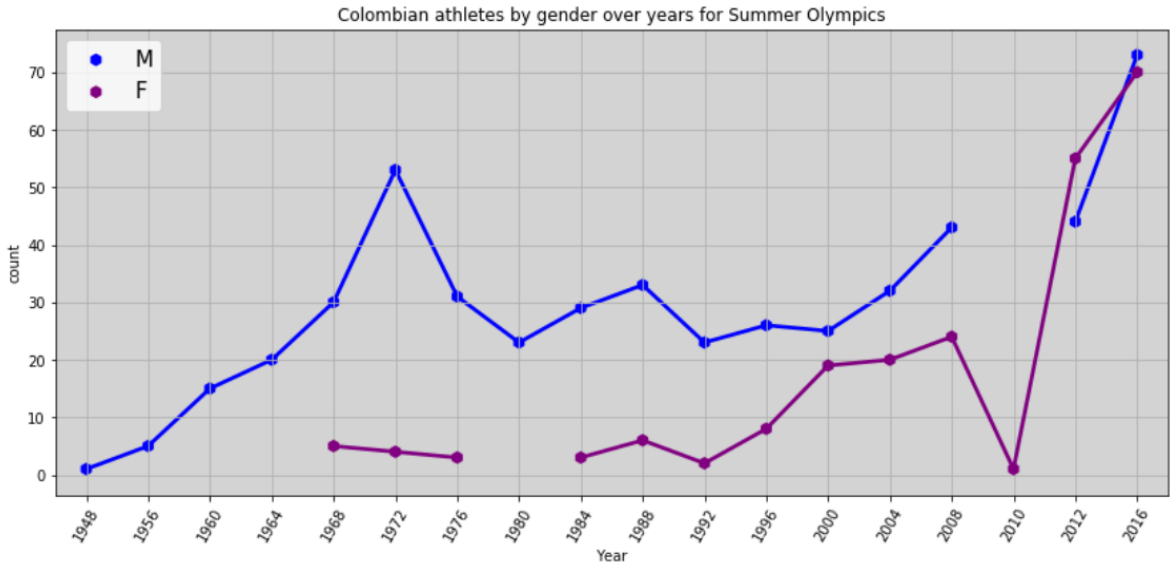
Info( ) del dataset de los atletas olímpicos Colombianos

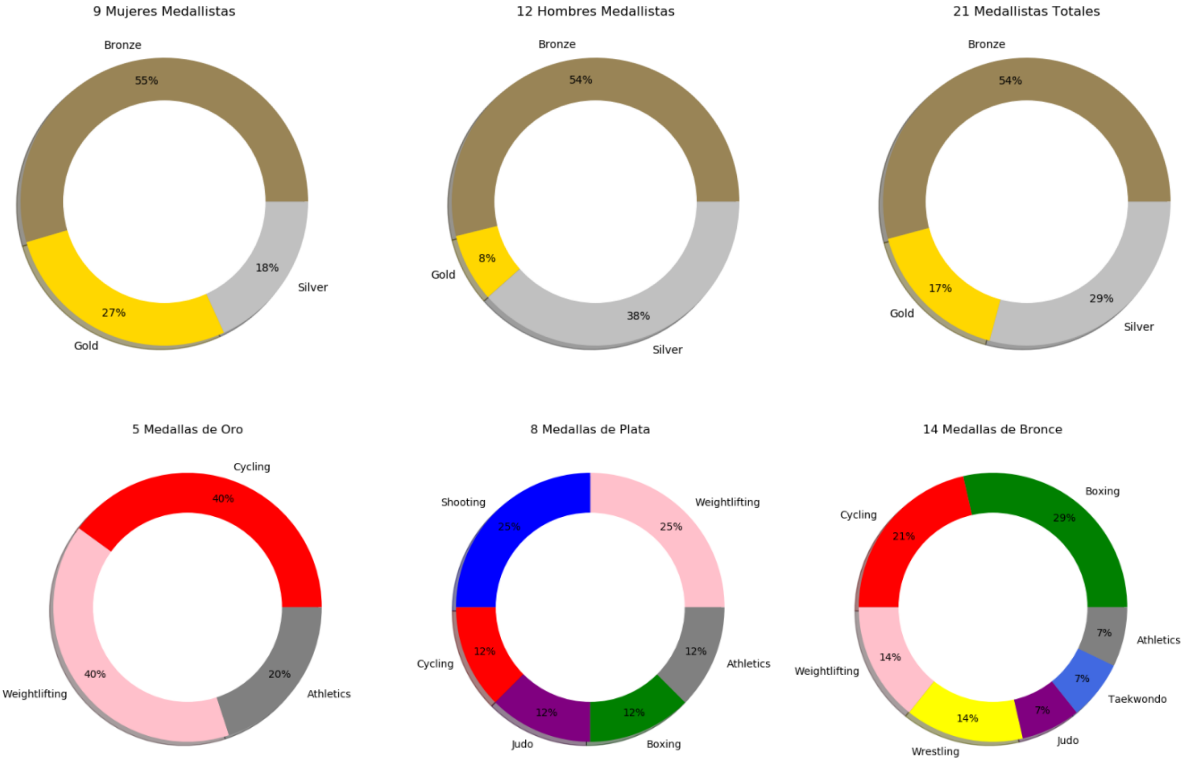
| | ID | Age (Years) | Height (cm) | Weight (kg) | Year |
|---|---|---|---|---|---|
| count | 945.000000 | 945.000000 | 945.000000 | 945.000000 | 945.000000 |
| mean | 63965.737566 | 24.985185 | 169.752381 | 64.868254 | 1994.205291 |
| std | 39962.490656 | 6.017963 | 9.047547 | 10.337798 | 18.654422 |
| min | 574.000000 | 12.000000 | 138.000000 | 38.000000 | 1948.000000 |
| 25% | 25877.000000 | 21.000000 | 164.000000 | 58.000000 | 1976.000000 |
| 50% | 74563.000000 | 24.000000 | 170.000000 | 64.000000 | 2000.000000 |
| 75% | 99964.000000 | 28.000000 | 175.000000 | 71.000000 | 2012.000000 |
| max | 135441.000000 | 52.000000 | 205.000000 | 110.000000 | 2016.000000 |

Colombia se encuentra en la posición 65 del ranking mundial de cantidad de medallas con 27 medallas, lo que representa el 0.0011 del total global. No es un desempeño alto, ya que la participación es relativamente baja con 945 atletas desde su primera participación en el 1948 y solo ha conseguido 27 medallas, es decir solo el 2.8% de participantes son medallistas (hasta el 2016 ).

**¿Qué deporte se debería incentivar más en el país para poder conseguir mejores resultados en los juegos olímpicos?**



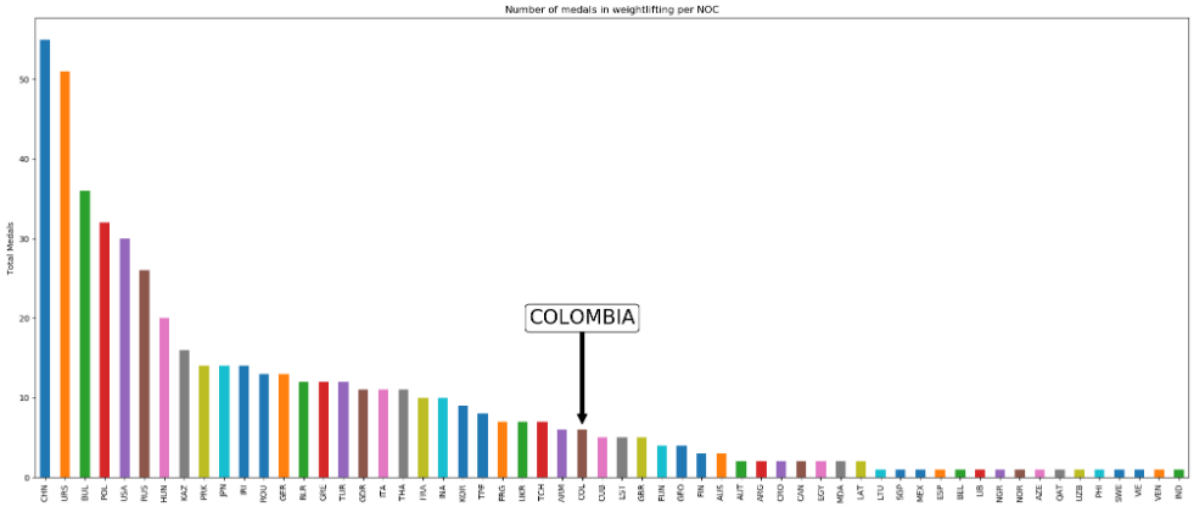


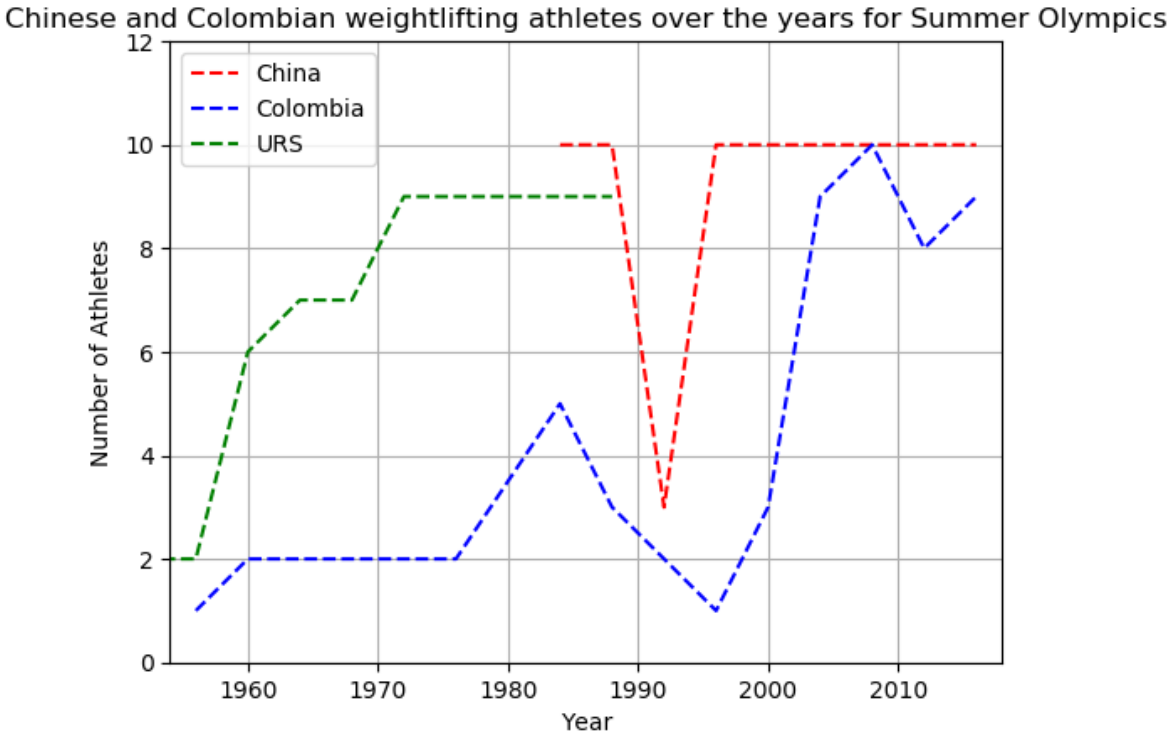




Ciclismo y Weightlifting tienen ambos 6 medallas, Weightlifting tiene 2 de plata, 2 de oro y 2 de bronce, comparado a ciclismo que tiene 1 de plata, 2 de Oro y 3 de bronce, además Weightlifting tiene casi un tercio de participantes de los que tiene ciclismo, es decir, con menor cantidad de participantes está ganando un mayor valor de medallas, lo que se traduce a una mayor tasa de éxito.

Como dato adicional, las mujeres aún con menor cantidad de participantes a través de los años ha ganado casi la misma cantidad de medallas que los hombres, pero de un mayor valor.

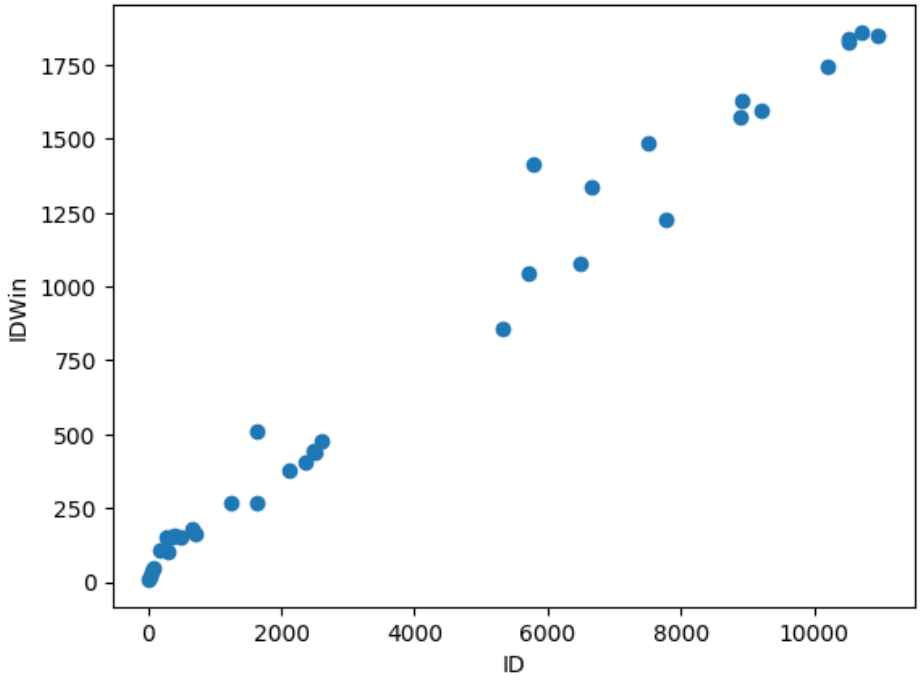
**¿Cómo ha sido el rendimiento de la delegación de weightlifting(levantamiento de pesas) colombiana a comparación de las potencias mundiales en dicho deporte?**

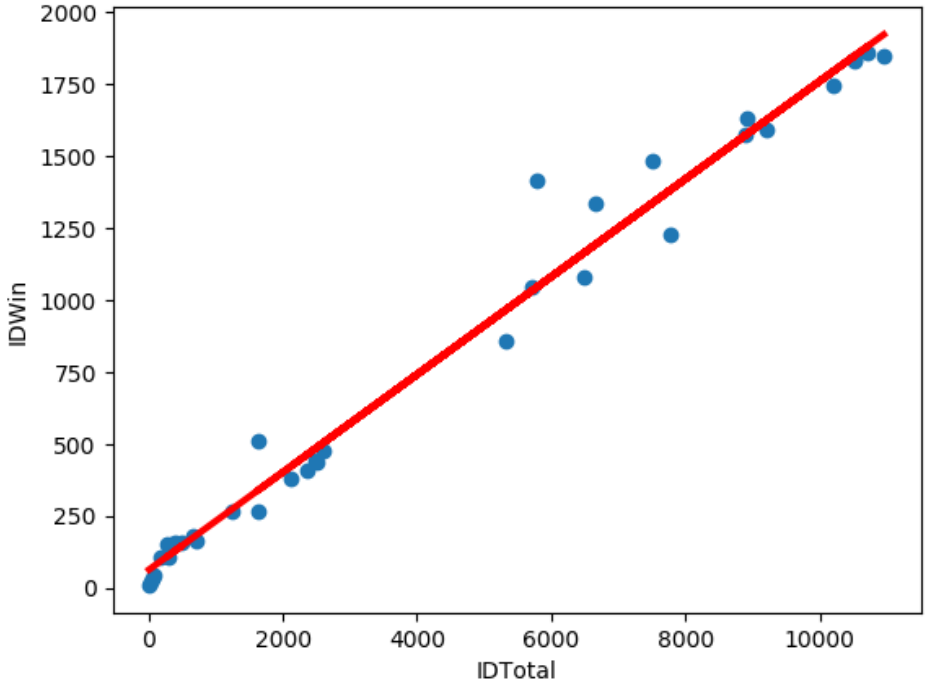



Mundialmente Colombia no está ubicado entre los mejores puestos, se encuentra en la posiciòn 28, aun así en la gràfica de cantidad de atletas a través de los años comparado con las potencias en Levantamiento de pesas se evidencia un incremento sustancial desde el 96, este crecimiento es proporcional al rendimiento en medallas, todas siendo entregadas a Colombia desde el año 2000 en adelante. Bronce (2004), Plata (2008,2008,2012), Oro (2000, 2016). Por lo tanto podemos argumentar que el Levantamiento de pesas es un deporte con gran rendimiento en los Olímpicos para Colombia.

**¿Dado N(cantidad determinada) atletas que compitan en representación de un NOC, cuantos de ellos podrían obtener una medalla de cualquier tipo?**




La regresión lineal predice una cantidad de medallistas de 145, con un score del 98%, es decir que el programa retorna un nivel de fiabilidad de casi el 100%, se obtiene un resultado satisfactorio de predicción con un score bastante alto, por lo cual se podría considerar la regresión lineal un modelo exitoso de predicción para esta pregunta de negocio.

**Realizado por: James Romero - Oscar Angarita - Andres Lindarte - Santiago Sandoval**