(https://databricks.com)

```
#convert data from csv to dataframe
path="---"
df_events = spark.read.format("csv").option("header", "true").load(path + "athlete_events.csv")
df_regions = spark.read.format("csv").option("header", "true").load(path + "noc_regions.csv")
#creating tables
df_events.write.mode("overwrite").saveAsTable("eventt")
df_regions.write.mode("overwrite").saveAsTable("region")
```

```
%sql
-- When importing data NA wasnt readed as NULL
UPDATE eventt
SET Age=NULL
WHERE Age='NA';

UPDATE eventt
SET Name=NULL
WHERE Name='NA';

UPDATE eventt
SET Sex=NULL
WHERE Sex='NA';

UPDATE eventt
SET Age=NULL
WHERE Age='NA';

UPDATE eventt
SET Height=NULL
WHERE Height='NA';

UPDATE eventt
SET Weight=NULL
WHERE Weight='NA';

UPDATE eventt
SET Team=NULL
WHERE Team='NA';

UPDATE eventt
SET NOC=NULL
WHERE NOC='NA';

UPDATE eventt
SET Games=NULL
WHERE Games='NA';

UPDATE eventt
SET Year=NULL
WHERE Year='NA';

UPDATE eventt
SET Season=NULL
WHERE Season='NA';

UPDATE eventt
SET City=NULL
WHERE City='NA';

UPDATE eventt
SET Sport=NULL
WHERE Sport='NA';

UPDATE eventt
SET Event=NULL
WHERE Event='NA';

UPDATE eventt
SET Medal=NULL
WHERE Medal='NA';
```

Table

New result table: ON

| | num_affected_rows |
|---|---|

| | |
|---|---|
| 1 | 231333 |

1 row

```
%fs rm -r dbfs:/user/hive/warehouse/TABLA
```

```
%sql
--Creating new table to change data types (park SQL doesn't directly
--support altering the data type of an existing column in a table)
CREATE TABLE IF NOT EXISTS sports_events (
  ID INT,
  Name STRING,
  Sex STRING,
  Age FLOAT,
  Height FLOAT,
  Weight FLOAT,
  Team STRING,
  NOC STRING,
  Games STRING,
  Year INT,
  Season STRING,
  City STRING,
  Sport STRING,
  Event STRING,
  Medal STRING
);
```

OK

```
%sql
INSERT INTO sports_events
SELECT
  ID,
  Name,
  Sex,
  CAST(Age AS FLOAT),     -- Convert Age column to FLOAT
  CAST(Height AS FLOAT),  -- Convert Height column to FLOAT
  CAST(Weight AS FLOAT),  -- Convert Weight column to FLOAT
  Team,
  NOC,
  Games,
  CAST(Year AS INT), -- Convert Year column to INT
  Season,
  City,
  Sport,
  Event,
  Medal
FROM eventt;
```

Table

New result table: ON

| | num_affected_rows | num_inserted_rows |
|---|---|---|
| 1 | 271116 | 271116 |

1 row

```
%sql
DROP TABLE IF EXISTS eventt;
```

OK

```
%sql
UPDATE sports_events
SET Name=TRIM('"' FROM Name);
```

Table

New result table: ON

| | num_affected_rows |
|---|---|
| 1 | 271116 |

1 row

Let's see how the data looks now

```
%sql
SELECT * FROM sports_events
```

Table

New result table: ON

| | ID | Name | Sex | Age | Height | Weight | Team | NOC | Games | Year |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | A Dijiang | M | 24 | 180 | 80 | China | CHN | 1992 Summer | 1992 |
| 2 | 2 | A Lamusi | M | 23 | 170 | 60 | China | CHN | 2012 Summer | 2012 |
| 3 | 3 | Gunnar Nielsen Aaby | M | 24 | null | null | Denmark | DEN | 1920 Summer | 1920 |
| 4 | 4 | Edgar Lindenau Aabye | M | 34 | null | null | Denmark/Sweden | DEN | 1900 Summer | 1900 |
| 5 | 5 | Christine Jacoba Aaftink | F | 21 | 185 | 82 | Netherlands | NED | 1988 Winter | 1988 |
| 6 | 5 | Christine Jacoba Aaftink | F | 21 | 185 | 82 | Netherlands | NED | 1988 Winter | 1988 |
| 7 | 5 | Christine Jacoba Aaftink | F | 25 | 185 | 82 | Netherlands | NED | 1992 Winter | 1992 |
| 8 | 5 | Christine Jacoba Aaftink | F | 25 | 185 | 82 | Netherlands | NED | 1992 Winter | 1992 |
| 9 | 5 | Christine Jacoba Aaftink | F | 27 | 185 | 82 | Netherlands | NED | 1994 Winter | 1994 |
| 10 | 5 | Christine Jacoba Aaftink | F | 27 | 185 | 82 | Netherlands | NED | 1994 Winter | 1994 |
| 11 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 12 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 13 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 14 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |

10,000+ rows | Truncated data

```
%sql
SELECT
    MAX(Age) max_age,
    MIN(Age) min_age,
    ROUND(AVG(Age),2) avg_age
FROM sports_events
```

Table

New result table: ON

| | max_age | min_age | avg_age |
|---|---|---|---|
| 1 | 97 | 10 | 25.56 |

1 row

```
%sql
SELECT Games, COUNT(ID) participants
FROM sports_events
GROUP BY Games
ORDER BY COUNT(ID) desc
```

Table | Visualization 1

New result table: ON

| | Games | participants |
|---|---|---|
| 1 | 2000 Summer | 13821 |
| 2 | 1996 Summer | 13780 |
| 3 | 2016 Summer | 13688 |
| 4 | 2008 Summer | 13602 |

| | | |
|---|---|---|
| 5 | 2004 Summer | 13443 |
| 6 | 1992 Summer | 12977 |
| 7 | 2012 Summer | 12920 |
| 8 | 1988 Summer | 12037 |
| 9 | 1972 Summer | 10304 |
| 10 | 1984 Summer | 9454 |
| 11 | 1976 Summer | 8641 |
| 12 | 1968 Summer | 8588 |
| 13 | 1952 Summer | 8270 |
| 14 | 1960 Summer | 8119 |
| 15 | 1964 Summer | 7702 |

51 rows

```
%sql
SELECT Name, COUNT(ID) GoldMedals
FROM sports_events
WHERE Medal='Gold'
GROUP BY Name
ORDER BY COUNT(ID) DESC
```

Table

New result table: ON

| | Name | GoldMedals |
|---|---|---|
| 1 | Michael Fred Phelps  II | 23 |
| 2 | Raymond Clarence ""Ray"" Ewry | 10 |
| 3 | Paavo Johannes Nurmi | 9 |
| 4 | Mark Andrew Spitz | 9 |
| 5 | Larysa Semenivna Latynina (Diriy-) | 9 |
| 6 | Frederick Carlton ""Carl"" Lewis | 9 |
| 7 | Usain St. Leo Bolt | 8 |
| 8 | Matthew Nicholas ""Matt"" Biondi | 8 |
| 9 | Ole Einar Bjrndalen | 8 |
| 10 | Birgit Fischer-Schmidt | 8 |
| 11 | Jennifer Elisabeth ""Jenny"" Thompson (-Cumpeli... | 8 |
| 12 | Sawao Kato | 8 |
| 13 | Vra slavsk (-Odloilov) | 7 |
| 14 | Nikolay Yefimovich Andrianov | 7 |
| 15 | Viktor Ivanovych Chukarin | 7 |

10,000+ rows | Truncated data

```sql
%sql
SELECT
  COALESCE(b.Team,gs._Team) Team,
  COALESCE(Gold_Medals,0)  GoldMedals,
  COALESCE(Silver_Medals,0) SilverMedals,
  COALESCE(Bronze_Medals,0) BronzeMedals,
  GoldMedals+SilverMedals+BronzeMedals Total
FROM
(
  SELECT Team, COUNT(ID) Bronze_Medals
  FROM sports_events
  WHERE Medal='Bronze'
  GROUP BY Team
  ORDER BY COUNT(ID) DESC
) AS b
FULL OUTER JOIN
(
  SELECT
    COALESCE(g.Team,S.Team) _Team,
    COALESCE(GoldMedals,0) Gold_Medals ,
    COALESCE(SilverMedals,0) Silver_Medals,
    Gold_Medals+Silver_Medals Total
  FROM
  (
    SELECT Team, COUNT(ID) GoldMedals
    FROM sports_events
    WHERE Medal='Gold'
    GROUP BY Team
    ORDER BY COUNT(ID) DESC
  ) g
  FULL OUTER JOIN
  (
    SELECT Team, COUNT(ID) SilverMedals
    FROM sports_events
    WHERE Medal='Silver'
    GROUP BY Team
    ORDER BY COUNT(ID) DESC
  ) s
  ON s.Team=g.Team
) gs
ON gs._Team=b.Team
```

Table

New result table: ON

| | Team | GoldMedals | SilverMedals | BronzeMedals | Total |
|---|---|---|---|---|---|
| 1 | "Large boat Central Naval Prep School ""Poros""-1" | 17 | 0 | 0 | 17 |
| 2 | "Life boat naval ship ""Spetsai""-1" | 0 | 7 | 0 | 7 |
| 3 | "Marinai della nave da guerra ""Varese""" | 7 | 0 | 17 | 24 |
| 4 | "Phalainis ton Thorichtou ""Hydra""-2" | 0 | 17 | 7 | 24 |
| 5 | A North American Team | 0 | 0 | 4 | 4 |
| 6 | Afghanistan | 0 | 0 | 2 | 2 |
| 7 | Algeria | 5 | 4 | 8 | 17 |
| 8 | Ali-Baba II | 0 | 0 | 5 | 5 |
| 9 | Amateur Athletic Association | 5 | 0 | 0 | 5 |
| 10 | Amstel Amsterdam | 0 | 0 | 4 | 4 |
| 11 | Ancora | 4 | 0 | 0 | 4 |
| 12 | Angelita | 12 | 0 | 0 | 12 |
| 13 | Antwerpia V | 0 | 0 | 5 | 5 |
| 14 | Aphrodite | 0 | 0 | 3 | 3 |
| 15 | Argentina | 91 | 84 | 91 | 266 |

498 rows

```sql
%sql
SELECT * FROM sports_events
```

Table

New result table: ON

| | ID | Name | Sex | Age | Height | Weight | Team | NOC | Games | Year |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | A Dijiang | M | 24 | 180 | 80 | China | CHN | 1992 Summer | 1992 |
| 2 | 2 | A Lamusi | M | 23 | 170 | 60 | China | CHN | 2012 Summer | 2012 |
| 3 | 3 | Gunnar Nielsen Aaby | M | 24 | null | null | Denmark | DEN | 1920 Summer | 1920 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 3 | Gunnar Nielsen Aaby | M | 24 | null | null | Denmark | DEN | 1920 Summer | 1920 |
| 4 | 4 | Edgar Lindenau Aabye | M | 34 | null | null | Denmark/Sweden | DEN | 1900 Summer | 1900 |
| 5 | 5 | Christine Jacoba Aaftink | F | 21 | 185 | 82 | Netherlands | NED | 1988 Winter | 1988 |
| 6 | 5 | Christine Jacoba Aaftink | F | 21 | 185 | 82 | Netherlands | NED | 1988 Winter | 1988 |
| 7 | 5 | Christine Jacoba Aaftink | F | 25 | 185 | 82 | Netherlands | NED | 1992 Winter | 1992 |
| 8 | 5 | Christine Jacoba Aaftink | F | 25 | 185 | 82 | Netherlands | NED | 1992 Winter | 1992 |
| 9 | 5 | Christine Jacoba Aaftink | F | 27 | 185 | 82 | Netherlands | NED | 1994 Winter | 1994 |
| 10 | 5 | Christine Jacoba Aaftink | F | 27 | 185 | 82 | Netherlands | NED | 1994 Winter | 1994 |
| 11 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 12 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 13 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 14 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |

10,000+ rows | Truncated data

```
%sql
SELECT
  COALESCE(winter.Year, summer.Year) year,
  COALESCE(WinterParticipants,0) Winter_Participant,
  COALESCE(SummerParticipants,0) Summer_Participant,
  Winter_Participant + Summer_Participant Total
FROM
(
  SELECT Year, COUNT(ID) WinterParticipants
  FROM sports_events
  WHERE Season='Winter'
  GROUP BY Year
  ORDER BY COUNT(ID) desc
) winter
FULL OUTER JOIN
(
  SELECT Year, COUNT(ID) SummerParticipants
  FROM sports_events
  WHERE Season='Summer'
  GROUP BY Year
  ORDER BY COUNT(ID) desc
) summer
ON winter.Year=summer.Year
```

Table | Visualization 1 | Visualization 2 | Visualization 3 — New result table: ON

| | year | Winter_Participant | Summer_Participant | Total |
|---|---|---|---|---|
| 1 | 1896 | 0 | 380 | 380 |
| 2 | 1900 | 0 | 1936 | 1936 |
| 3 | 1904 | 0 | 1301 | 1301 |
| 4 | 1906 | 0 | 1733 | 1733 |
| 5 | 1908 | 0 | 3101 | 3101 |
| 6 | 1912 | 0 | 4040 | 4040 |
| 7 | 1920 | 0 | 4292 | 4292 |
| 8 | 1924 | 460 | 5233 | 5693 |
| 9 | 1928 | 582 | 4992 | 5574 |
| 10 | 1932 | 352 | 2969 | 3321 |
| 11 | 1936 | 895 | 6506 | 7401 |
| 12 | 1948 | 1075 | 6405 | 7480 |
| 13 | 1952 | 1088 | 8270 | 9358 |
| 14 | 1956 | 1307 | 5127 | 6434 |
| 15 | 1960 | 1116 | 8119 | 9235 |

35 rows

```
%sql
SELECT Name, NOC, COUNT(ID) GoldMedals
FROM sports_events
WHERE Medal='Gold'
GROUP BY Name, NOC
ORDER BY COUNT(ID) DESC
LIMIT 20
```

Table — New result table: ON

| | Name | NOC | GoldMedals |
|---|---|---|---|
| 1 | Michael Fred Phelps II | USA | 23 |
| 2 | Raymond Clarence ""Ray"" Ewry | USA | 10 |
| 3 | Paavo Johannes Nurmi | FIN | 9 |
| 4 | Mark Andrew Spitz | USA | 9 |
| 5 | Larysa Semenivna Latynina (Diriy-) | URS | 9 |
| 6 | Frederick Carlton ""Carl"" Lewis | USA | 9 |
| 7 | Matthew Nicholas ""Matt"" Biondi | USA | 8 |
| 8 | Ole Einar Bjrndalen | NOR | 8 |
| 9 | Usain St. Leo Bolt | JAM | 8 |
| 10 | Jennifer Elisabeth ""Jenny"" Thompson (-Cumpeli... | USA | 8 |
| 11 | Sawao Kato | JPN | 8 |
| 12 | Vra slavsk (-Odloilov) | TCH | 7 |
| 13 | Nikolay Yefimovich Andrianov | URS | 7 |
| 14 | Viktor Ivanovych Chukarin | URS | 7 |
| 15 | Borys Anfiyanovych Shakhlin | URS | 7 |

20 rows

```
%sql
SELECT
  NOC,
  COUNT(ID) GoldMedals,
  ROW_NUMBER() OVER (ORDER BY COUNT(ID) DESC) AS CountryPosition
FROM sports_events
WHERE Medal='Gold'
GROUP BY NOC
ORDER BY GoldMedals DESC
```

Table

New result table: ON

| | NOC | GoldMedals | CountryPosition |
|---|---|---|---|
| 1 | USA | 2638 | 1 |
| 2 | URS | 1082 | 2 |
| 3 | GER | 745 | 3 |
| 4 | GBR | 678 | 4 |
| 5 | ITA | 575 | 5 |
| 6 | FRA | 501 | 6 |
| 7 | SWE | 479 | 7 |
| 8 | CAN | 463 | 8 |
| 9 | HUN | 432 | 9 |
| 10 | GDR | 397 | 10 |
| 11 | RUS | 390 | 11 |
| 12 | NOR | 378 | 12 |
| 13 | CHN | 350 | 13 |
| 14 | AUS | 348 | 14 |
| 15 | NED | 287 | 15 |

109 rows

```sql
%sql
--In the top 20 competitors with the most Gold Medals the average country position in total gold
--medals is 7.8, this means that the top competitors with the most gold medals are from the
--countries which got more gold medals over all.
SELECT AVG(CountryPosition) AvgPos_Top20
FROM
(
  SELECT
    Top20.Name,
    Top20.GoldMedals,
    Top20.NOC,
    Pos.CountryPosition
  FROM
  (
    SELECT Name, NOC, COUNT(ID) GoldMedals
    FROM sports_events
    WHERE Medal='Gold'
    GROUP BY Name, NOC
    ORDER BY COUNT(ID) DESC
    LIMIT 20
  ) Top20
  LEFT OUTER JOIN
  (
    SELECT
      NOC,
      COUNT(ID) GoldMedals,
      ROW_NUMBER() OVER (ORDER BY COUNT(ID) DESC) AS CountryPosition
    FROM sports_events
    WHERE Medal='Gold'
    GROUP BY NOC
    ORDER BY GoldMedals DESC
  ) Pos
  ON Top20.NOC=Pos.NOC
)
```

Table

New result table: ON

| | 1.2 AvgPos_Top20 |
|---|---|
| 1 | 7.8 |

1 row

```sql
%sql
SELECT SUM(Winners) AS `Not top sport total winners`
FROM
(
  SELECT
    Sport,
    COUNT(ID) Winners
  FROM sports_events
  WHERE NOC='USA' AND Medal='Gold'
  GROUP BY Sport
)
WHERE
  Sport!='Swimming' AND
  Sport!='Athletics' AND
  Sport!='Basketball' AND
  Sport!='Rowing'
```

Table

New result table: ON

| | 1²3 Not top sport total winners |
|---|---|
| 1 | 980 |

1 row

```
%sql
SELECT
  SUM(Winners) `Top sports total winners`
FROM
(
  SELECT
    Sport,
    COUNT(ID) Winners
  FROM sports_events
  WHERE NOC='USA' AND Medal='Gold'
  GROUP BY Sport
)
WHERE
  Sport='Swimming' OR
  Sport='Athletics' OR
  Sport='Basketball' OR
  Sport='Rowing'
```

Table

New result table: ON

| | Top sports total winners |
|---|---|
| 1 | 1658 |

1 row

1658 gold medals were achieved by Swimming, Athletics, Basketball, Rowing

980 gold medals were achieved without those 4 disciplines (the others are 40)

Therefore, 62% of the gold medals were achieved by the 9% of the medal-winning disciplines of USA

```
%sql
SELECT
  NOC,
  COUNT(ID) GoldMedals,
  ROW_NUMBER() OVER (ORDER BY COUNT(ID) DESC) AS CountryPosition
FROM sports_events
WHERE Medal='Gold'
GROUP BY NOC
ORDER BY GoldMedals DESC
```

Table

New result table: ON

| | NOC | GoldMedals | CountryPosition |
|---|---|---|---|
| 1 | USA | 2638 | 1 |
| 2 | URS | 1082 | 2 |
| 3 | GER | 745 | 3 |
| 4 | GBR | 678 | 4 |
| 5 | ITA | 575 | 5 |
| 6 | FRA | 501 | 6 |
| 7 | SWE | 479 | 7 |
| 8 | CAN | 463 | 8 |
| 9 | HUN | 432 | 9 |
| 10 | GDR | 397 | 10 |
| 11 | RUS | 390 | 11 |
| 12 | NOR | 378 | 12 |
| 13 | CHN | 350 | 13 |
| 14 | AUS | 348 | 14 |
| 15 | NED | 287 | 15 |

109 rows

```
%sql
SELECT
  COUNT(Top100.PersonalPosition) AS total_top100,
  Pos.CountryPosition
FROM
(
  SELECT
    ROW_NUMBER() OVER (ORDER BY COUNT(ID) DESC) AS PersonalPosition,
    Name,
    NOC,
    COUNT(ID) GoldMedals
  FROM sports_events
  WHERE Medal='Gold'
  GROUP BY Name, NOC
  LIMIT 100
) Top100
LEFT OUTER JOIN
(
  SELECT
    NOC,
    COUNT(ID) GoldMedals,
    ROW_NUMBER() OVER (ORDER BY COUNT(ID) DESC) AS CountryPosition
  FROM sports_events
  WHERE Medal='Gold'
  GROUP BY NOC
  ORDER BY GoldMedals DESC
) Pos
ON Top100.NOC=Pos.NOC
GROUP BY Pos.CountryPosition
ORDER BY total_top100 DESC
```

Table | Visualization 1 — New result table: ON

| | total_top100 | CountryPosition |
|---|---|---|
| 1 | 32 | 1 |
| 2 | 10 | 2 |
| 3 | 7 | 4 |
| 4 | 6 | 9 |
| 5 | 6 | 3 |
| 6 | 5 | 12 |
| 7 | 5 | 16 |
| 8 | 4 | 13 |
| 9 | 3 | 18 |
| 10 | 3 | 5 |
| 11 | 3 | 11 |
| 12 | 3 | 22 |
| 13 | 2 | 20 |
| 14 | 2 | 14 |
| 15 | 1 | 42 |

23 rows

The amount of top-ranked personal winners grows as the position of the team. This means top 1 team tend to have more competitors in the top100 than the top 2 team and so

```
%sql

SELECT
  COUNT(ID) MedalWinners,
  Year
FROM sports_events
WHERE
  Medal IS NOT NULL AND
  Season='Summer' AND
  Year>1936
GROUP BY Year
```

Table | Visualization 1 — New result table: ON

| | MedalWinners | Year |
|---|---|---|
| 1 | 897 | 1952 |

| | | |
|---|---|---|
| 2 | 893 | 1956 |
| 3 | 1582 | 1988 |
| 4 | 1057 | 1968 |
| 5 | 2001 | 2004 |
| 6 | 1842 | 1996 |
| 7 | 911 | 1960 |
| 8 | 1941 | 2012 |
| 9 | 2023 | 2016 |
| 10 | 1384 | 1980 |
| 11 | 1215 | 1972 |
| 12 | 1712 | 1992 |
| 13 | 1476 | 1984 |
| 14 | 2004 | 2000 |
| 15 | 1029 | 1964 |

18 rows

Number of medal winners increments by year. The reason of this is clearly that the amounts of competitors and sports also grew.

```
from pyspark.sql import SparkSession

# Creating a SparkSession
spark = SparkSession.builder \
    .appName("Convert SQL table to Python lists") \
    .getOrCreate()

# Reading the SQL table into a DataFrame
df = spark.sql("""
 SELECT
    COUNT(Top100.PersonalPosition) AS total_top100,
    CountryPosition
  FROM
  (
    SELECT
      ROW_NUMBER() OVER (ORDER BY COUNT(ID) DESC) AS PersonalPosition,
      Name,
      NOC,
      COUNT(ID) GoldMedals
    FROM sports_events
    WHERE Medal='Gold'
    GROUP BY Name, NOC
    LIMIT 100
  ) Top100
  LEFT OUTER JOIN
  (
    SELECT
      NOC,
      COUNT(ID) GoldMedals,
      ROW_NUMBER() OVER (ORDER BY COUNT(ID) DESC) AS CountryPosition
    FROM sports_events
    WHERE Medal='Gold'
    GROUP BY NOC
    ORDER BY GoldMedals DESC
  ) Pos
  ON Top100.NOC=Pos.NOC
  GROUP BY Pos.CountryPosition
  ORDER BY CountryPosition ASC
""")
```

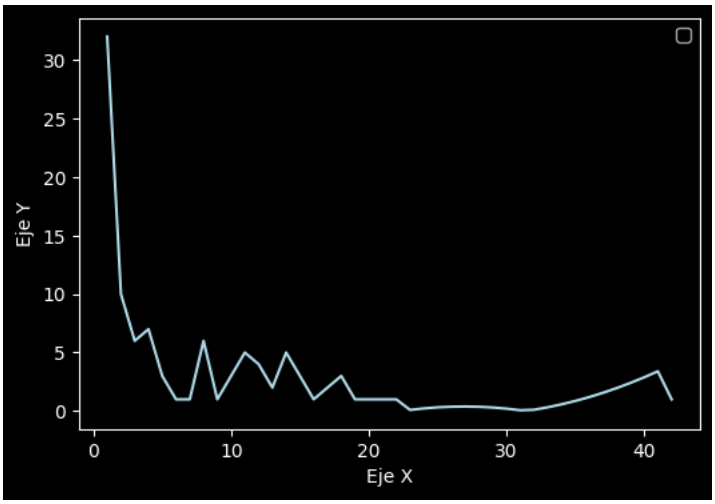
```
# Converting the DataFrame columns into Python lists
total_top100 = df.select("total_top100").rdd.flatMap(lambda x: x).collect()
country_position = df.select("CountryPosition").rdd.flatMap(lambda x: x).collect()
```

```
import numpy as np
import matplotlib.pyplot as plt
```

```
#Fill mising country position values
coefs=np.polyfit(country_position,total_top100,2)
resize_country_position=np.arange(1, 43, 1, dtype=int)
resize_total_top100=np.abs(np.polyval(coefs, resize_country_position))
for x in range(len(total_top100)):
    resize_total_top100[x-1]=total_top100[x-1]
```

```
plt.style.use("dark_background")
fig, ax = plt.subplots(figsize=(6,4))
plt.plot(resize_country_position, resize_total_top100,color="lightblue")
ax.set_xlabel("Eje X")
ax.set_ylabel("Eje Y")
ax.legend()
plt.show()
```

```
No artists with labels found to put in legend.  Note that artists whose label start with an underscore are ignored when legend() is called with no argument.
```




```
%sql
--Not medal winners
SELECT
  ROUND(AVG(CASE WHEN Sex='M' THEN Height END),2) MaleAvgHeight,
  ROUND(AVG(CASE WHEN Sex='F' THEN Height END),2) FemaleAvgHeight
FROM(
  SELECT
    ID,
    Height,
    Sex,
    SUM(MedalBinary)
  FROM(
    SELECT
      ID,
      Height,
      Sex,
      CASE WHEN Medal IS NOT NULL THEN 1 ELSE 0 END MedalBinary
    FROM sports_events
    WHERE Year>2000
  )
  GROUP BY ID, Height, Sex
  HAVING SUM(MedalBinary)=0
)
```

Table

New result table: ON

| | 1.2 MaleAvgHeight | 1.2 FemaleAvgHeight |
|---|---|---|
| 1 | 181.21 | 169.05 |

1 row

```
%sql
--Medal winners
SELECT
  ROUND(AVG(CASE WHEN Sex='M' THEN Height END),2) MaleAvgHeight,
  ROUND(AVG(CASE WHEN Sex='F' THEN Height END),2) FemaleAvgHeight
FROM(
  SELECT
    ID,
    Height,
    Sex,
    SUM(MedalBinary)
  FROM(
    SELECT
      ID,
      Height,
      Sex,
      CASE WHEN Medal IS NOT NULL THEN 1 ELSE 0 END MedalBinary
    FROM sports_events
    WHERE Year>2000
  )
  GROUP BY ID, Height, Sex
  HAVING SUM(MedalBinary)>0
)
```

Table

New result table: ON

| | 1.2 MaleAvgHeight | 1.2 FemaleAvgHeight |
|---|---|---|
| 1 | 183.18 | 171.07 |

1 row

New Metric

```
%sql
SELECT
  *,
  CASE WHEN Medal IS NOT NULL THEN 1 ELSE 0 END MedalBinary
FROM sports_events
```

Table

New result table: ON

| | 123 ID | ABC Name | ABC Sex | 1.2 Age | 1.2 Height | 1.2 Weight | ABC Team | ABC NOC | ABC Games | 123 Year |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | A Dijiang | M | 24 | 180 | 80 | China | CHN | 1992 Summer | 1992 |
| 2 | 2 | A Lamusi | M | 23 | 170 | 60 | China | CHN | 2012 Summer | 2012 |
| 3 | 3 | Gunnar Nielsen Aaby | M | 24 | null | null | Denmark | DEN | 1920 Summer | 1920 |
| 4 | 4 | Edgar Lindenau Aabye | M | 34 | null | null | Denmark/Sweden | DEN | 1900 Summer | 1900 |
| 5 | 5 | Christine Jacoba Aaftink | F | 21 | 185 | 82 | Netherlands | NED | 1988 Winter | 1988 |
| 6 | 5 | Christine Jacoba Aaftink | F | 21 | 185 | 82 | Netherlands | NED | 1988 Winter | 1988 |
| 7 | 5 | Christine Jacoba Aaftink | F | 25 | 185 | 82 | Netherlands | NED | 1992 Winter | 1992 |
| 8 | 5 | Christine Jacoba Aaftink | F | 25 | 185 | 82 | Netherlands | NED | 1992 Winter | 1992 |
| 9 | 5 | Christine Jacoba Aaftink | F | 27 | 185 | 82 | Netherlands | NED | 1994 Winter | 1994 |
| 10 | 5 | Christine Jacoba Aaftink | F | 27 | 185 | 82 | Netherlands | NED | 1994 Winter | 1994 |
| 11 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 12 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 13 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 14 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |

10,000+ rows | Truncated data

Table

New result table: ON

| | 123 ID | ABC Name | ABC Sex | 1.2 Age | 1.2 Height | 1.2 Weight | ABC Team | ABC NOC | ABC Games | 123 Year |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | A Dijiang | M | 24 | 180 | 80 | China | CHN | 1992 Summer | 1992 |
| 2 | 2 | A Lamusi | M | 23 | 170 | 60 | China | CHN | 2012 Summer | 2012 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 3 | Gunnar Nielsen Aaby | M | 24 | null | null | Denmark | DEN | 1920 Summer | 1920 |
| 4 | 4 | Edgar Lindenau Aabye | M | 34 | null | null | Denmark/Sweden | DEN | 1900 Summer | 1900 |
| 5 | 5 | Christine Jacoba Aaftink | F | 21 | 185 | 82 | Netherlands | NED | 1988 Winter | 1988 |
| 6 | 5 | Christine Jacoba Aaftink | F | 21 | 185 | 82 | Netherlands | NED | 1988 Winter | 1988 |
| 7 | 5 | Christine Jacoba Aaftink | F | 25 | 185 | 82 | Netherlands | NED | 1992 Winter | 1992 |
| 8 | 5 | Christine Jacoba Aaftink | F | 25 | 185 | 82 | Netherlands | NED | 1992 Winter | 1992 |
| 9 | 5 | Christine Jacoba Aaftink | F | 27 | 185 | 82 | Netherlands | NED | 1994 Winter | 1994 |
| 10 | 5 | Christine Jacoba Aaftink | F | 27 | 185 | 82 | Netherlands | NED | 1994 Winter | 1994 |
| 11 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 12 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 13 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |
| 14 | 6 | Per Knut Aaland | M | 31 | 188 | 75 | United States | USA | 1992 Winter | 1992 |

10,000+ rows | Truncated data