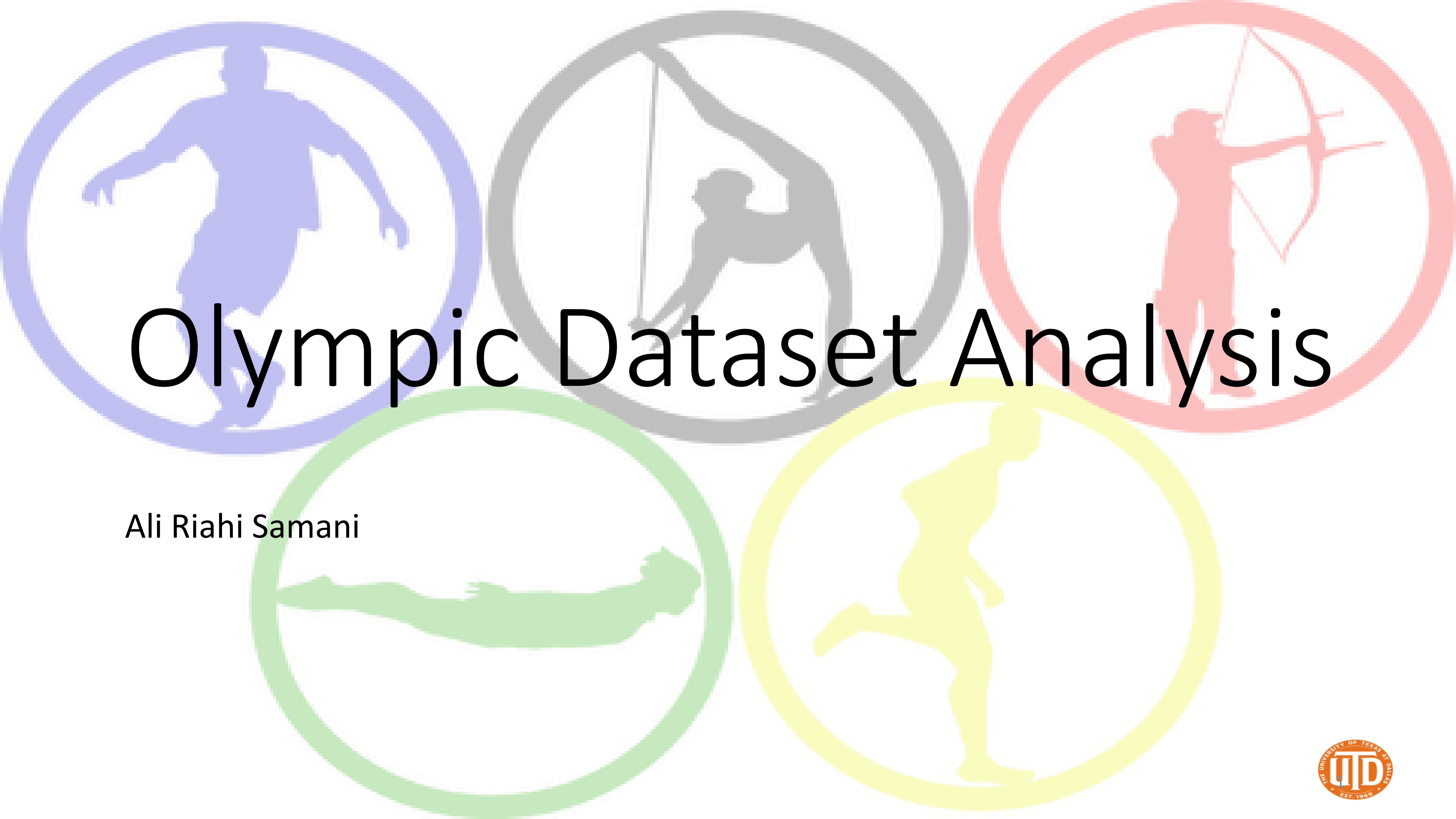


Olympic Dataset Analysis

Ali Riahi Samani

Olympics and Dataset

- The first ancient Olympic Games can be traced back to 776 BC.
- They were dedicated to the Olympian gods and were staged on the ancient plains of Olympia.
- The dataset contains details of 120 years of modern Olympic Games from Athens 1896 to Rio 2016.

- This dataset provides an opportunity to ask questions about how the Olympics have evolved over time, including questions about the participation and performance of women, different nations, and different sports and events.

Tools: Hadoop, Hive

Performed Analytics

- Top 5 countries who have won the maximum number of medals over the years and the corresponding male versus female distribution of athletes (medal winners) for these countries?

- For the USA, in the last 5 Olympics which sports have brought maximum gold, silver and bronze medals respectively?

- What are the average age, height, and weight of a male or a female winning a gold, silver, and bronze respectively in badminton?

- Top and Bottom 5 Olympics(years) in terms of the number of countries participated and the count of male and female athletes in those years?

- Top 5 sports in terms of how many times they were part of an Olympics over the years, and which were the years they were NOT part of an Olympics?

- Top 5 countries that had the biggest difference between their summer and winter gold medal counts?

- Top 5 cities which have hosted most number of Olympics in summer and winter respectively?

Top 5 Countries Winning most Medals

Top 5 countries who have won the maximum number of medals over the years and the corresponding male versus female distribution of athletes (medal winners) for these countries.

Columns used: id, sport, event, noc, region, medal, sex

Winning Sports in USA in Last 5 Olympics

For the USA, in the last 5 Olympics which sports have brought maximum gold, silver and bronze medals

Columns used: id, sport, event, year, medal, noc, region

Badminton

What are the average physical characteristics of a male or a female winning a gold, silver, and bronze respectively in badminton?

Columns Used: id, sport, event, medal, age, height, weight, sex

Highest and Lowest Participation Years

Top and Bottom 5 Olympics(years) in terms of the number of countries participated and the count of male and female athletes in those years

Columns used: id, sport, event, noc, region, year, sex

Top 5 Sports in Olympics

Top 5 sports in terms of how many times they were part of an Olympics over the years, and which were the years they were NOT part of an Olympics

Columns used: id, sport, event, year

Top 5 Country with Max Summer - Winter Diff

Top 5 countries that had the biggest difference between their summer and winter gold medal counts

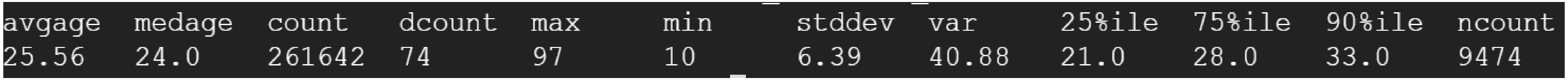
Columns used: id, sport, event, noc, region, year, season, medal

Top 5 Host Cities in Summer and Winter

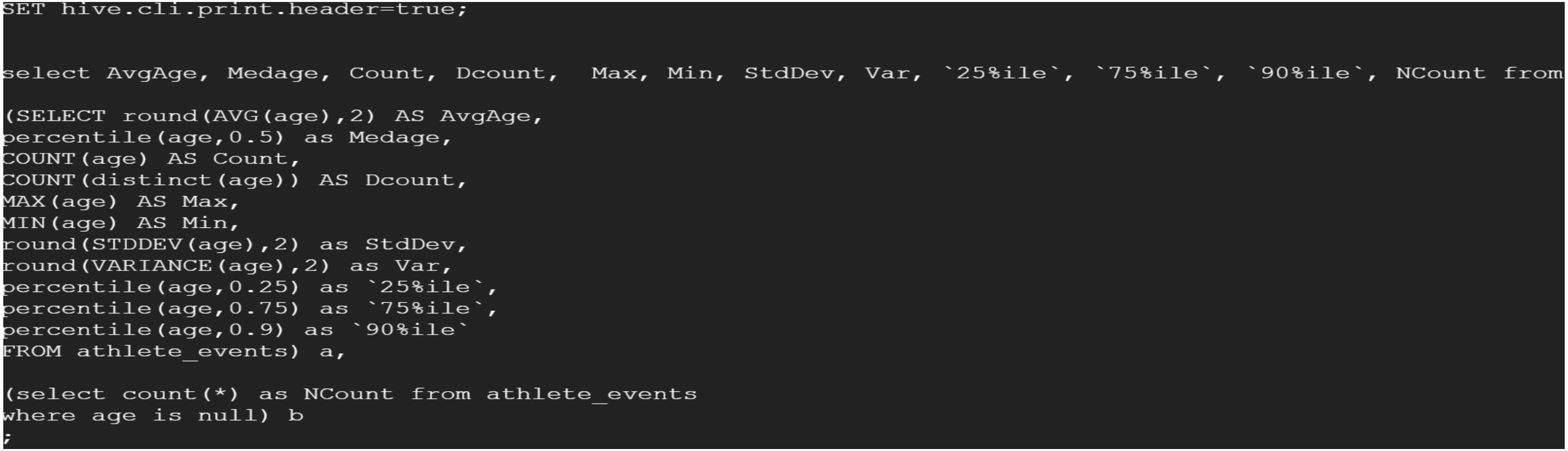
Top 5 cities which have hosted most number of Olympics in summer and winter respectively

Columns used: id, sport, event, city, season, year, noc, region

Column Analytics : Age

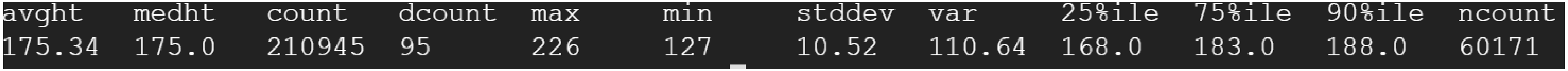


```
avgage  medage  count   dcount  max     min     stddev  var     25%ile  75%ile  90%ile  ncount
25.56   24.0    261642  74      97      10      6.39    40.88   21.0    28.0    33.0    9474
```

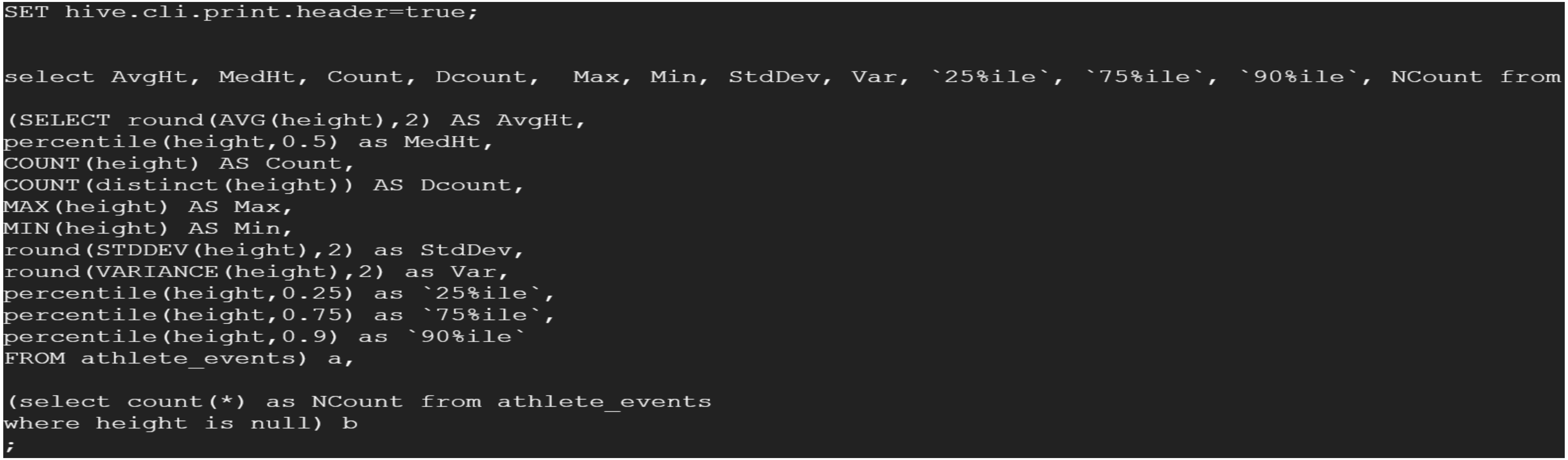


```
SET hive.cli.print.header=true;


select AvgAge, Medage, Count, Dcount,  Max, Min, StdDev, Var, `25%ile`, `75%ile`, `90%ile`, NCount from

(SELECT round(AVG(age),2) AS AvgAge,
percentile(age,0.5) as Medage,
COUNT(age) AS Count,
COUNT(distinct(age)) AS Dcount,
MAX(age) AS Max,
MIN(age) AS Min,
round(STDDEV(age),2) as StdDev,
round(VARIANCE(age),2) as Var,
percentile(age,0.25) as `25%ile`,
percentile(age,0.75) as `75%ile`,
percentile(age,0.9) as `90%ile`
FROM athlete_events) a,

(select count(*) as NCount from athlete_events
where age is null) b
;
```

Column Analytics : Height

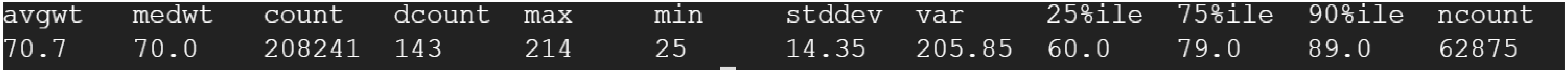


```
avght    medht    count    dcount   max      min      stddev   var      25%ile   75%ile   90%ile   ncount
175.34   175.0    210945   95       226      127      10.52    110.64   168.0    183.0    188.0    60171
```

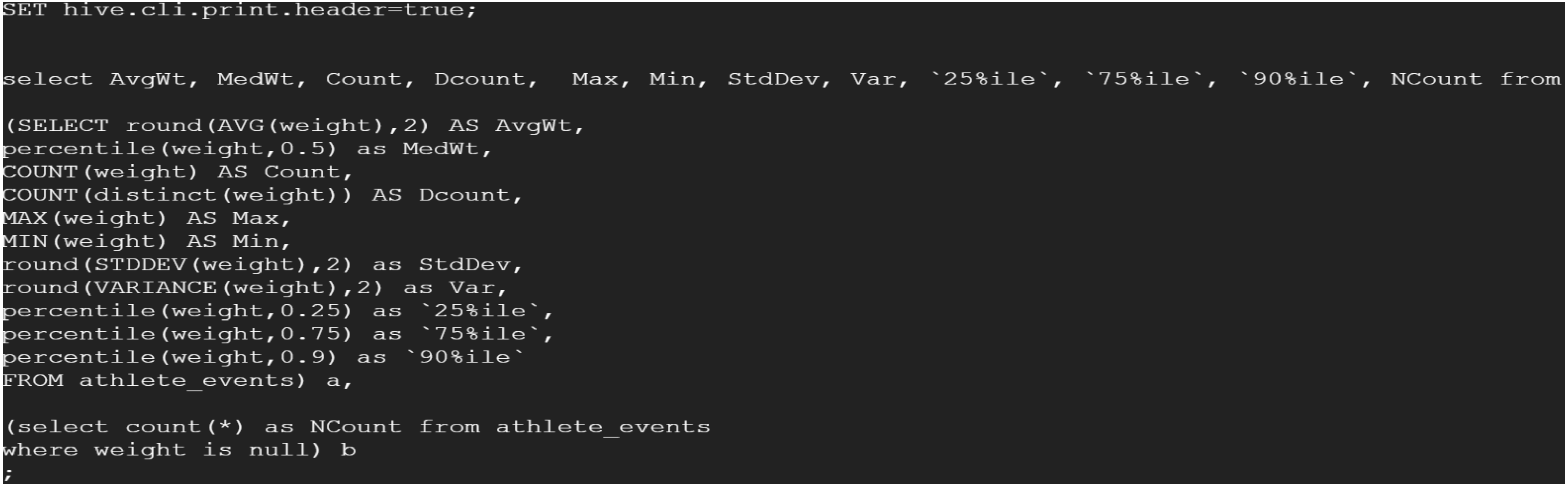


```
SET hive.cli.print.header=true;


select AvgHt, MedHt, Count, Dcount,  Max, Min, StdDev, Var, `25%ile`, `75%ile`, `90%ile`, NCount from

(SELECT round(AVG(height),2) AS AvgHt,
percentile(height,0.5) as MedHt,
COUNT(height) AS Count,
COUNT(distinct(height)) AS Dcount,
MAX(height) AS Max,
MIN(height) AS Min,
round(STDDEV(height),2) as StdDev,
round(VARIANCE(height),2) as Var,
percentile(height,0.25) as `25%ile`,
percentile(height,0.75) as `75%ile`,
percentile(height,0.9) as `90%ile`
FROM athlete_events) a,

(select count(*) as NCount from athlete_events
where height is null) b
;
```

Column Analytics : Weight

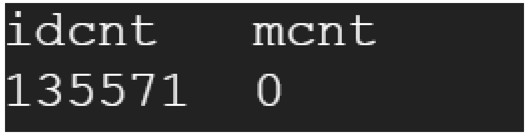


```
avgwt   medwt   count   dcount  max     min     stddev  var     25%ile  75%ile  90%ile  ncount
70.7    70.0    208241  143     214     25      14.35   205.85  60.0    79.0    89.0    62875
```

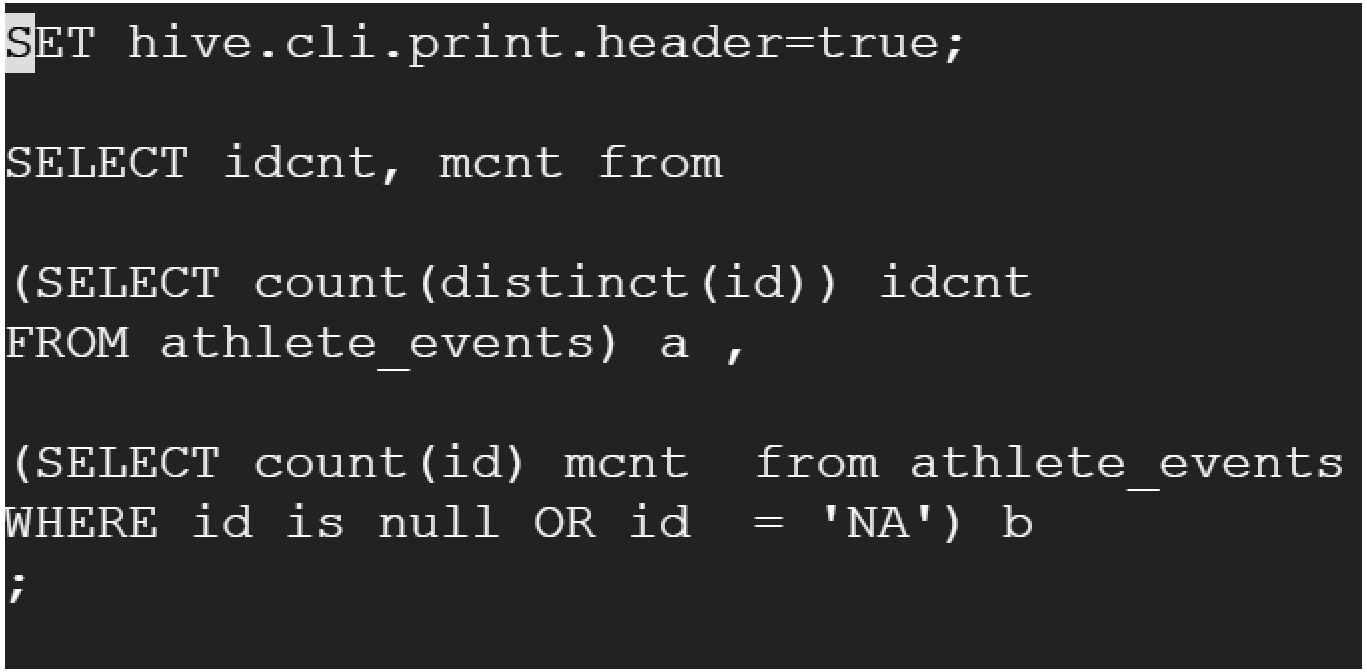


```
SET hive.cli.print.header=true;


select AvgWt, MedWt, Count, Dcount,  Max, Min, StdDev, Var, `25%ile`, `75%ile`, `90%ile`, NCount from

(SELECT round(AVG(weight),2) AS AvgWt,
percentile(weight,0.5) as MedWt,
COUNT(weight) AS Count,
COUNT(distinct(weight)) AS Dcount,
MAX(weight) AS Max,
MIN(weight) AS Min,
round(STDDEV(weight),2) as StdDev,
round(VARIANCE(weight),2) as Var,
percentile(weight,0.25) as `25%ile`,
percentile(weight,0.75) as `75%ile`,
percentile(weight,0.9) as `90%ile`
FROM athlete_events) a,

(select count(*) as NCount from athlete_events
where weight is null) b
;
```

Column Analytics : id

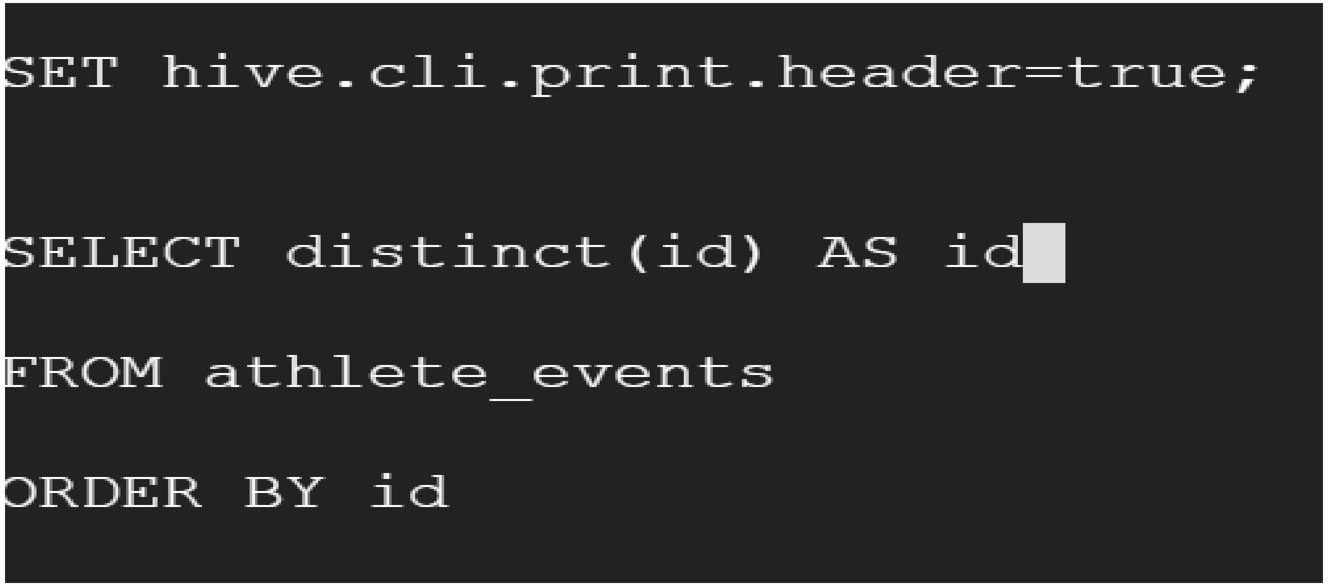


```
idcnt    mcnt
135571   0
```



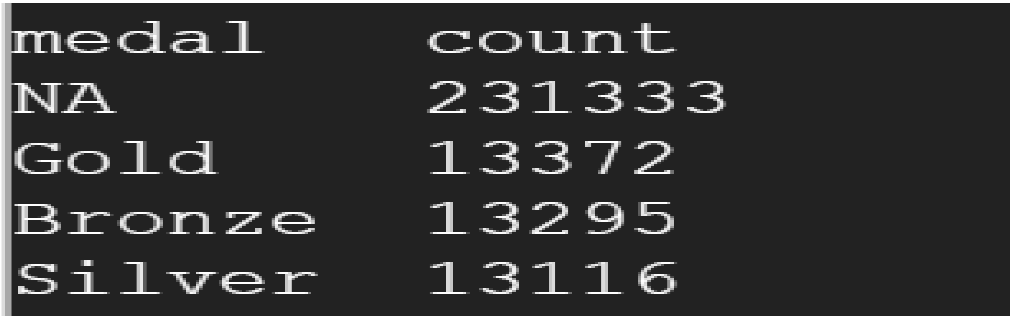
```
SET hive.cli.print.header=true;

SELECT idcnt, mcnt from

(SELECT count(distinct(id)) idcnt
FROM athlete_events) a ,

(SELECT count(id) mcnt  from athlete_events
WHERE id is null OR id  = 'NA') b
;
```

Column Analytics : id

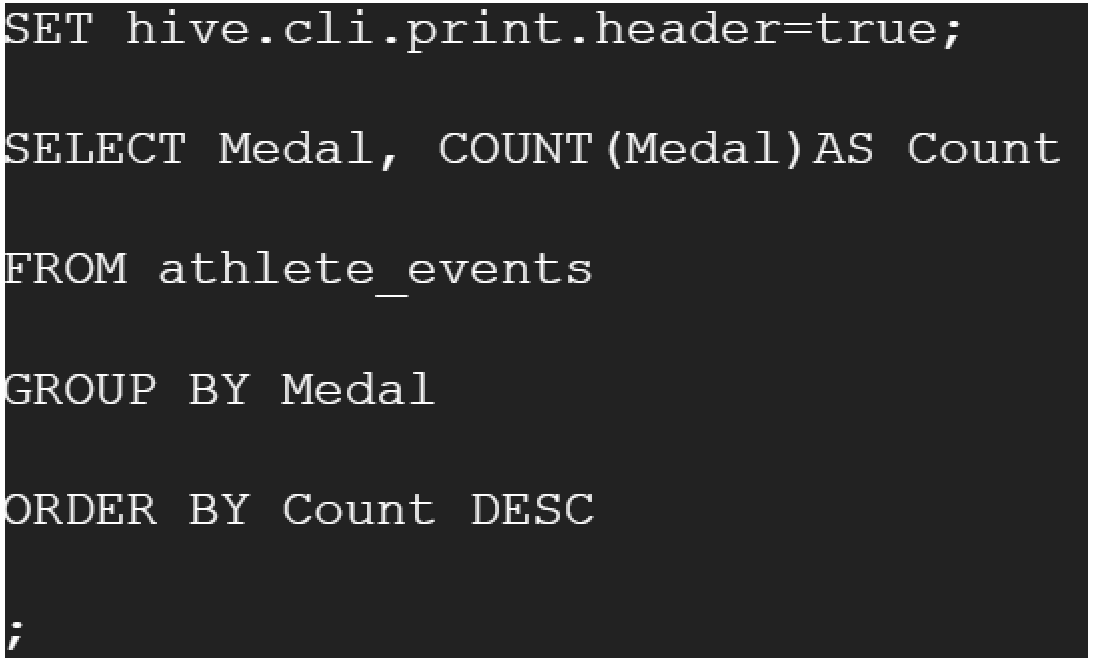


```
SET hive.cli.print.header=true;

SELECT distinct(id) AS id
FROM athlete_events
ORDER BY id
```

Column Analytics : Medals

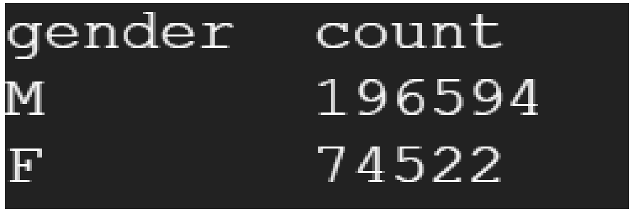


```
medal     count
NA        231333
Gold      13372
Bronze    13295
Silver    13116
```

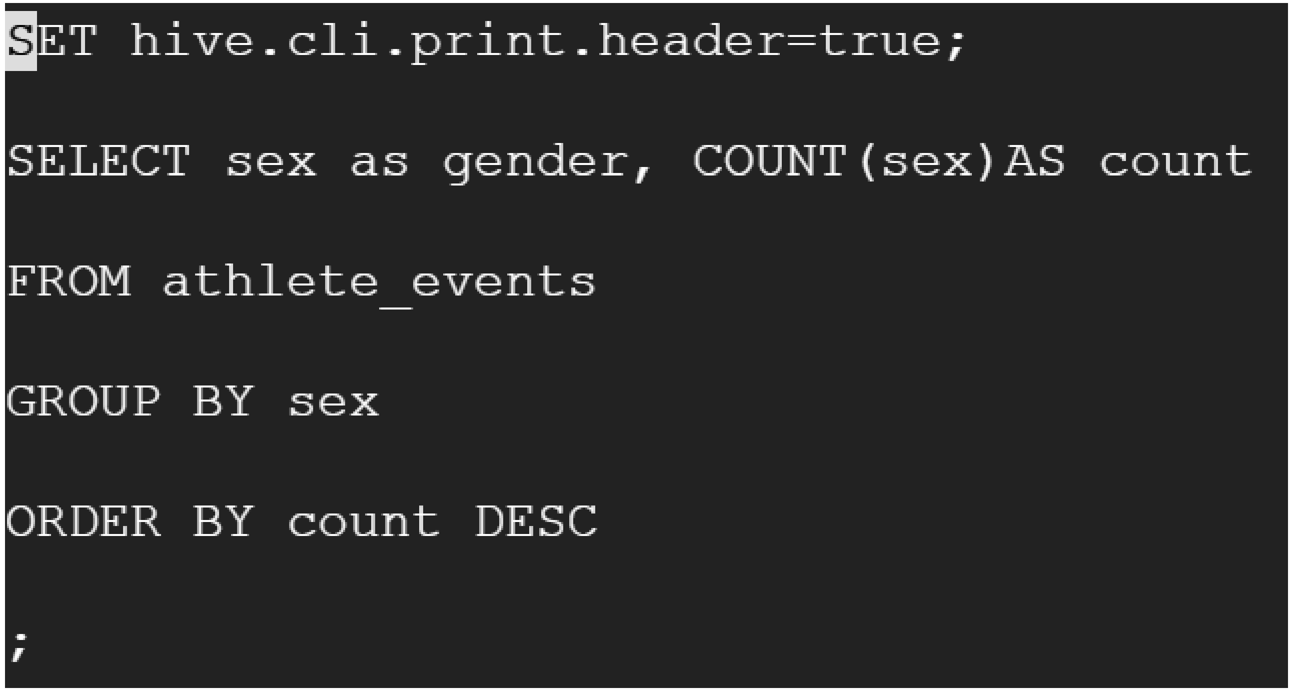


```
SET hive.cli.print.header=true;

SELECT Medal, COUNT(Medal)AS Count

FROM athlete_events

GROUP BY Medal

ORDER BY Count DESC

;
```

Column Analytics : Sex

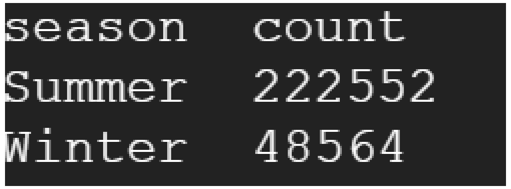


```
gender  count
M       196594
F       74522
```

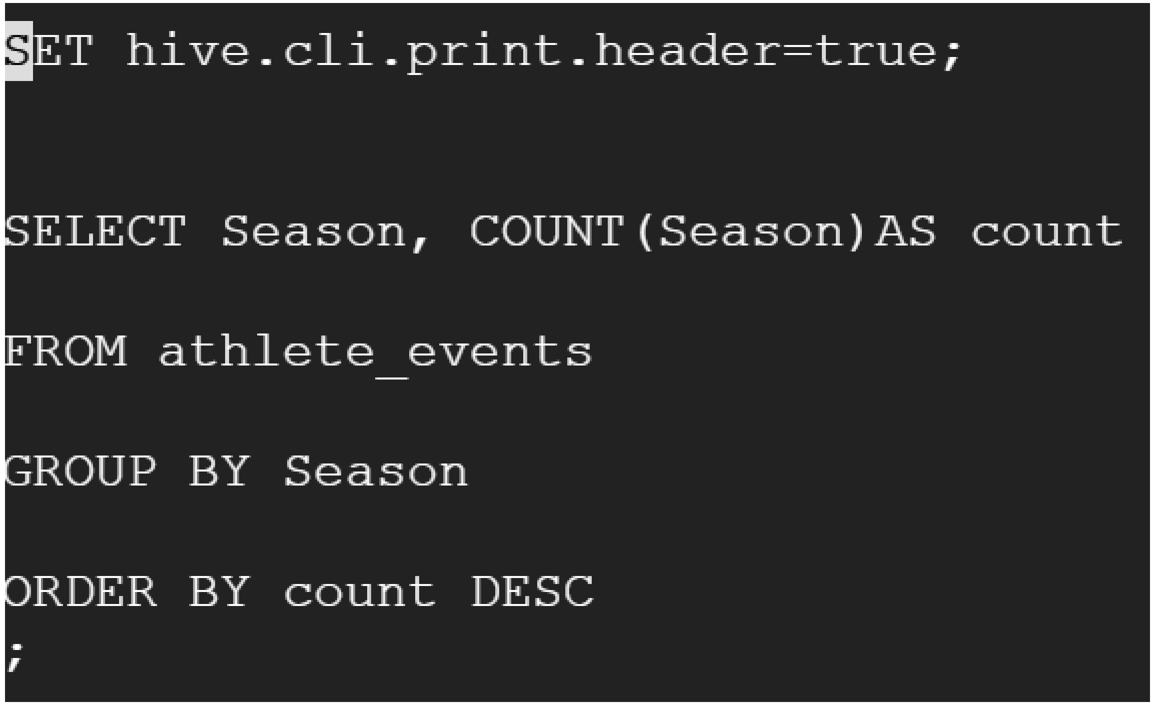


```
SET hive.cli.print.header=true;

SELECT sex as gender, COUNT(sex)AS count

FROM athlete_events

GROUP BY sex

ORDER BY count DESC

;
```

Column Analytics : Season

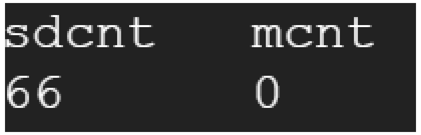


```
season   count
Summer   222552
Winter   48564
```

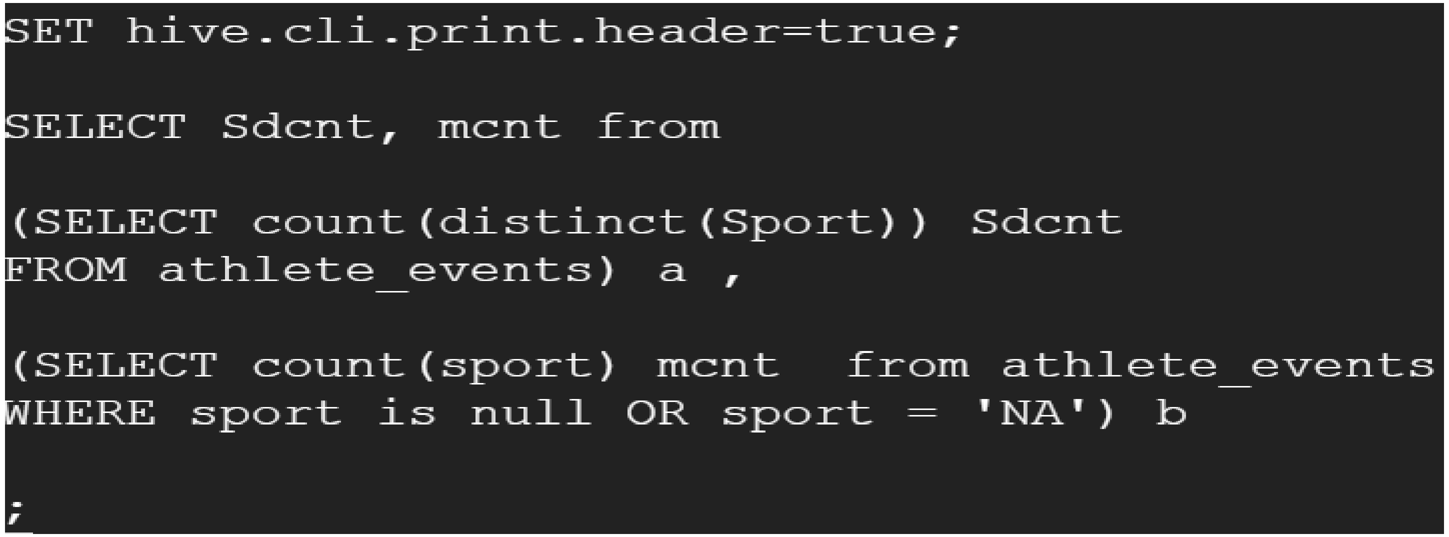


```
SET hive.cli.print.header=true;


SELECT Season, COUNT(Season)AS count

FROM athlete_events

GROUP BY Season

ORDER BY count DESC
;
```

Column Analytics : Sport

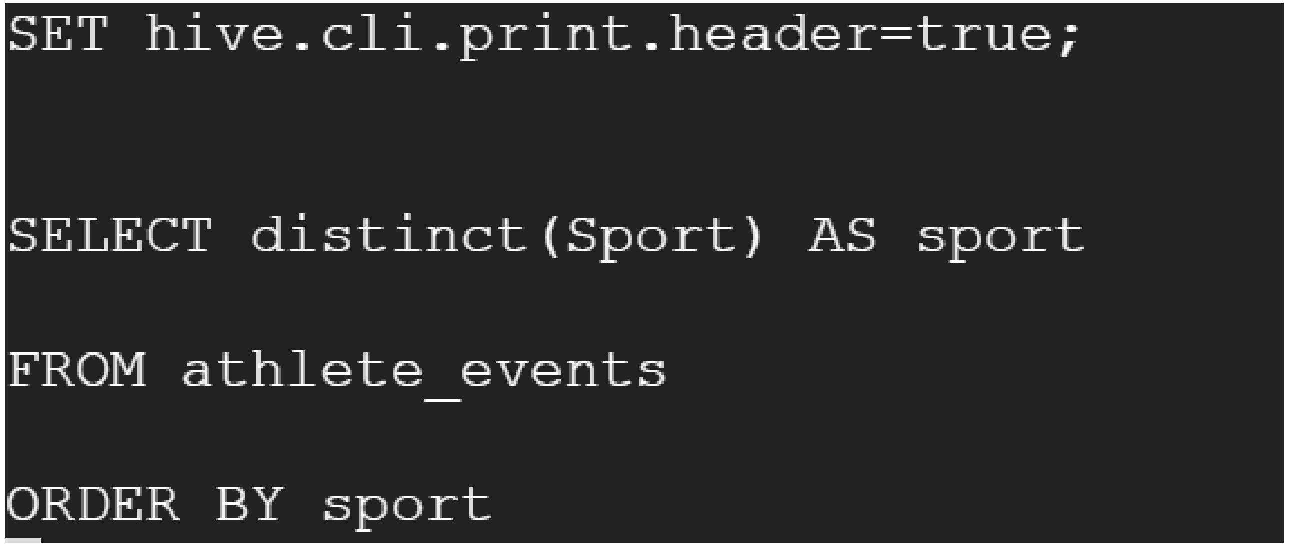


```
sdcnt      mcnt
66         0
```



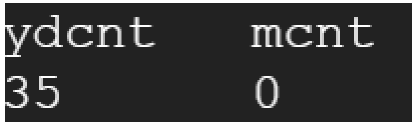
```
SET hive.cli.print.header=true;

SELECT Sdcnt, mcnt from

(SELECT count(distinct(Sport)) Sdcnt
FROM athlete_events) a ,

(SELECT count(sport) mcnt  from athlete_events
WHERE sport is null OR sport = 'NA') b

;
```

Column Analytics : Sport

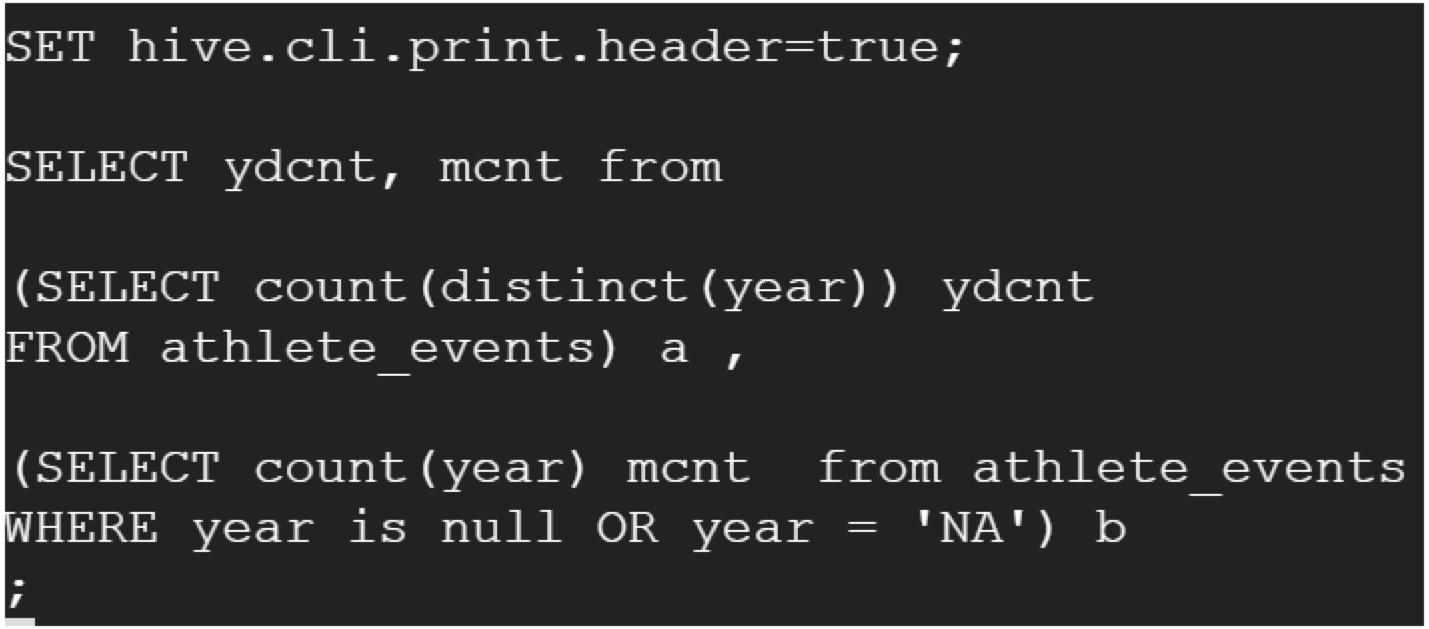


```
SET hive.cli.print.header=true;


SELECT distinct(Sport) AS sport

FROM athlete_events

ORDER BY sport
```

Column Analytics : Year

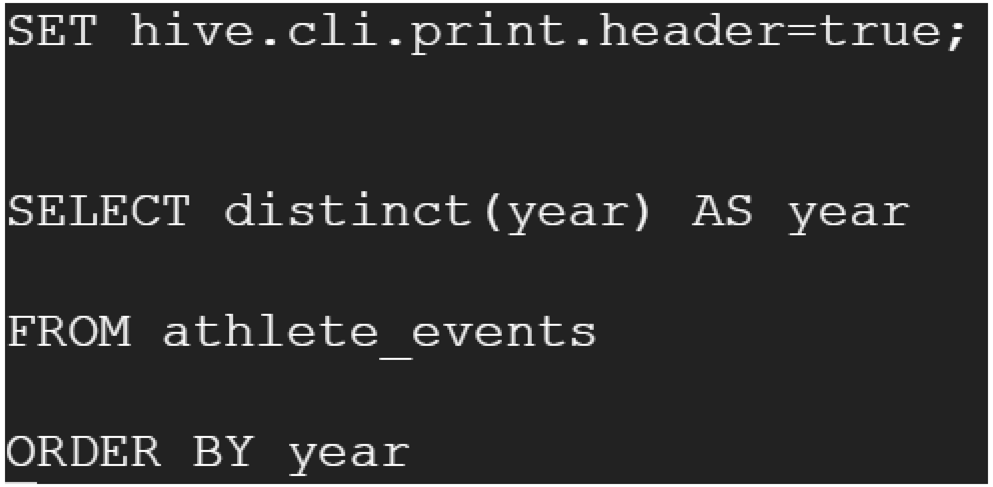


```
ydcnt     mcnt
35        0
```

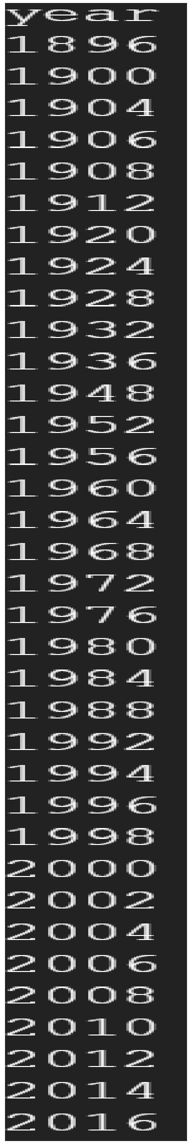


```
SET hive.cli.print.header=true;

SELECT ydcnt, mcnt from

(SELECT count(distinct(year)) ydcnt
FROM athlete_events) a ,

(SELECT count(year) mcnt  from athlete_events
WHERE year is null OR year = 'NA') b
;
```

Column Analytics : Year

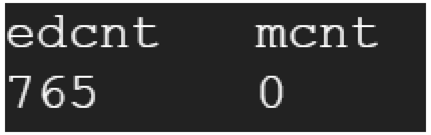


```
SET hive.cli.print.header=true;


SELECT distinct(year) AS year

FROM athlete_events

ORDER BY year
```

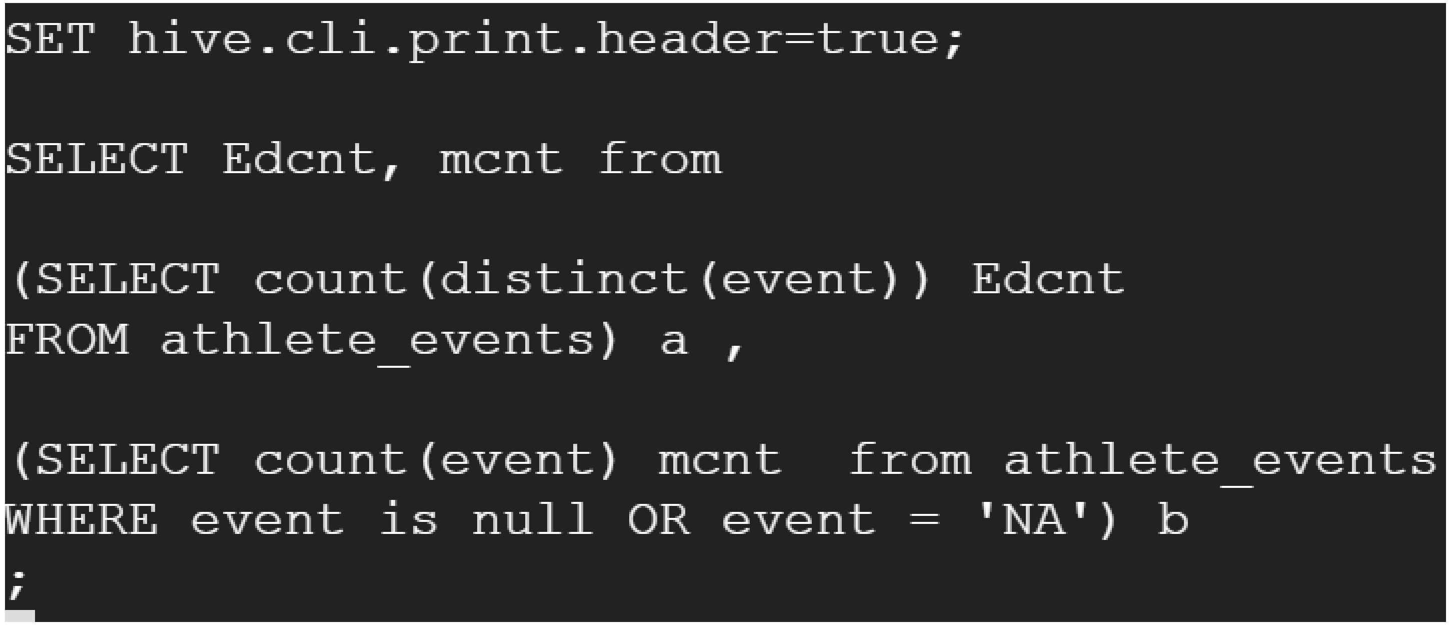


```
year
1896
1900
1904
1906
1908
1912
1920
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
```

Column Analytics : Event

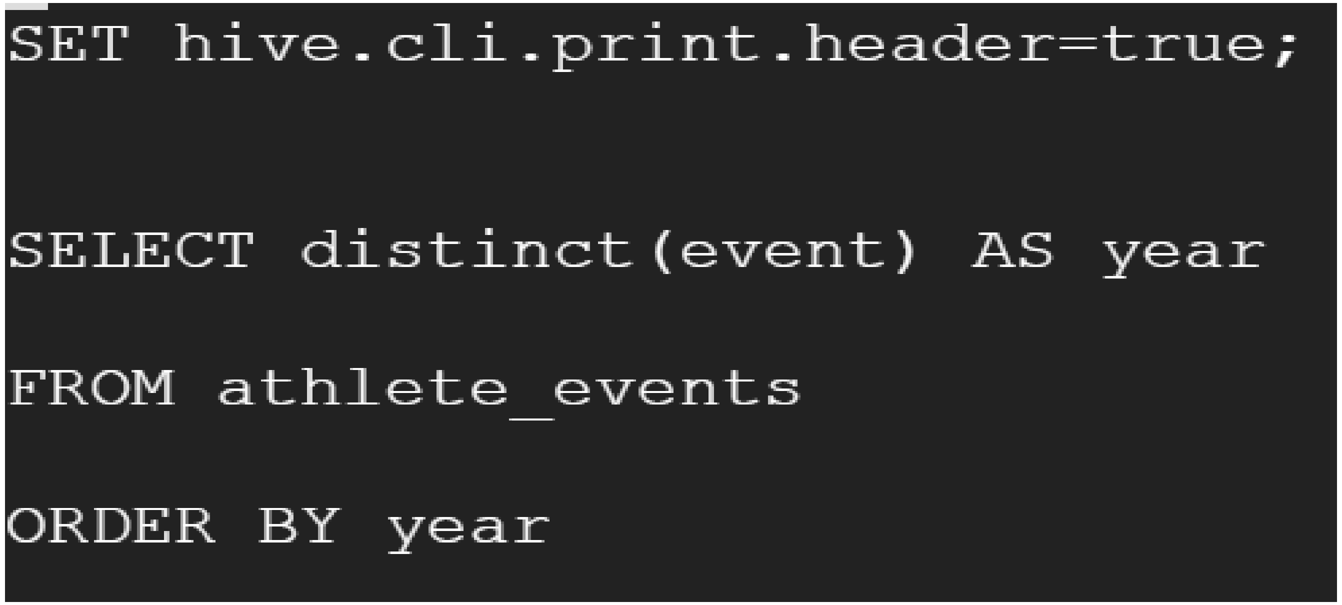


```
edcnt     mcnt
765       0
```



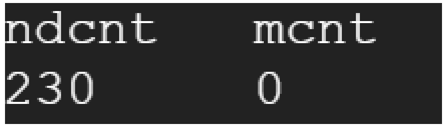
```
SET hive.cli.print.header=true;

SELECT Edcnt, mcnt from

(SELECT count(distinct(event)) Edcnt
FROM athlete_events) a ,

(SELECT count(event) mcnt  from athlete_events
WHERE event is null OR event = 'NA') b
;
```

Column Analytics : Event

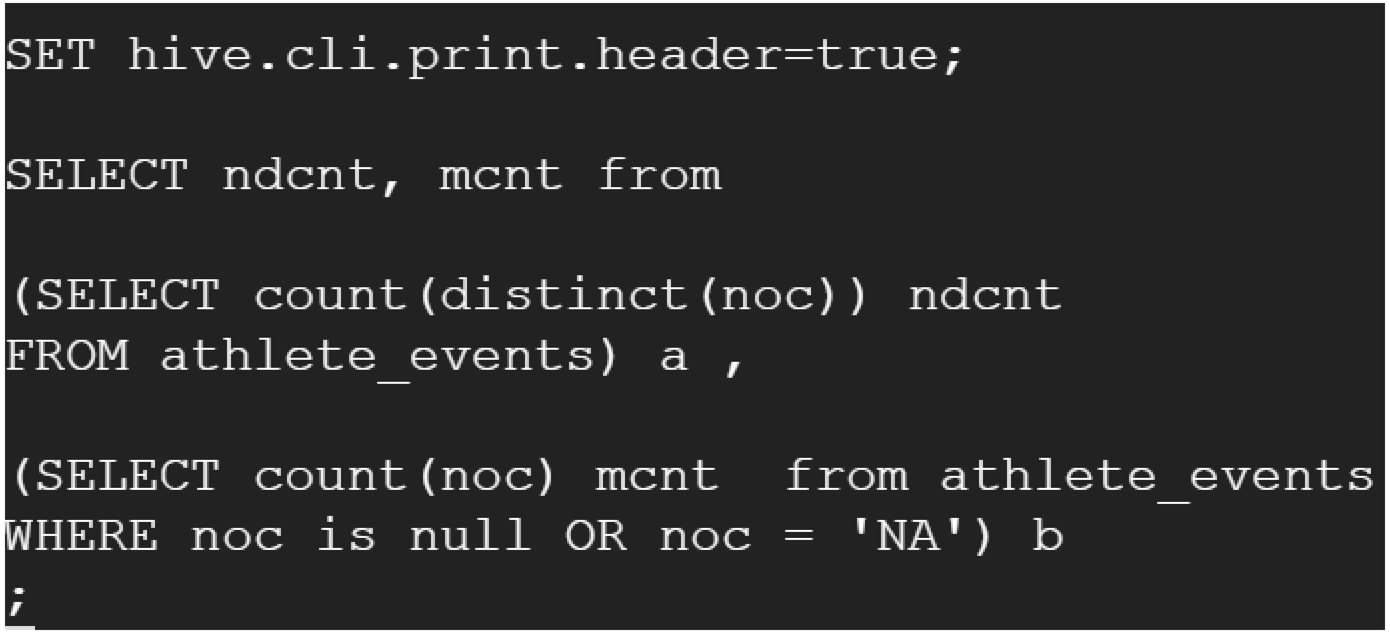


```
SET hive.cli.print.header=true;


SELECT distinct(event) AS year

FROM athlete_events

ORDER BY year
```

Column Analytics : NOC

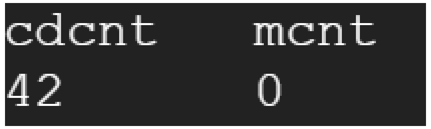


```
ndcnt     mcnt
230       0
```

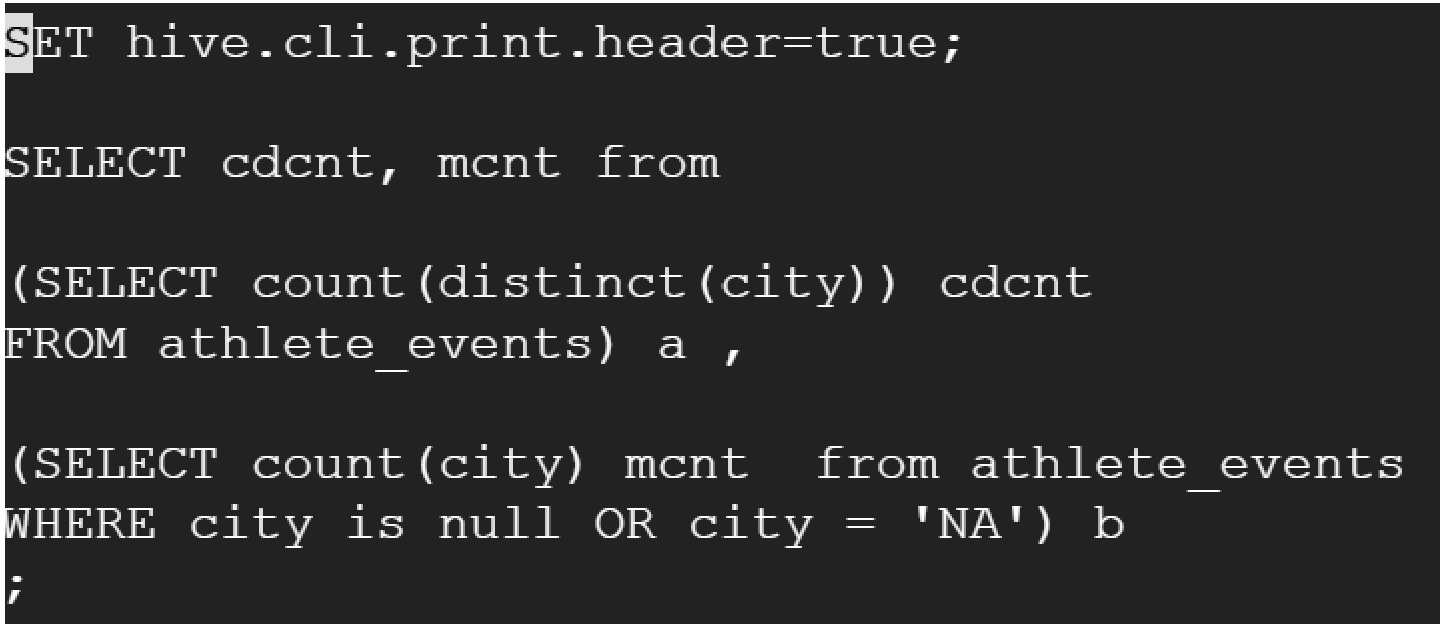


```
SET hive.cli.print.header=true;

SELECT ndcnt, mcnt from

(SELECT count(distinct(noc)) ndcnt
FROM athlete_events) a ,

(SELECT count(noc) mcnt  from athlete_events
WHERE noc is null OR noc = 'NA') b
;
```

Column Analytics : City

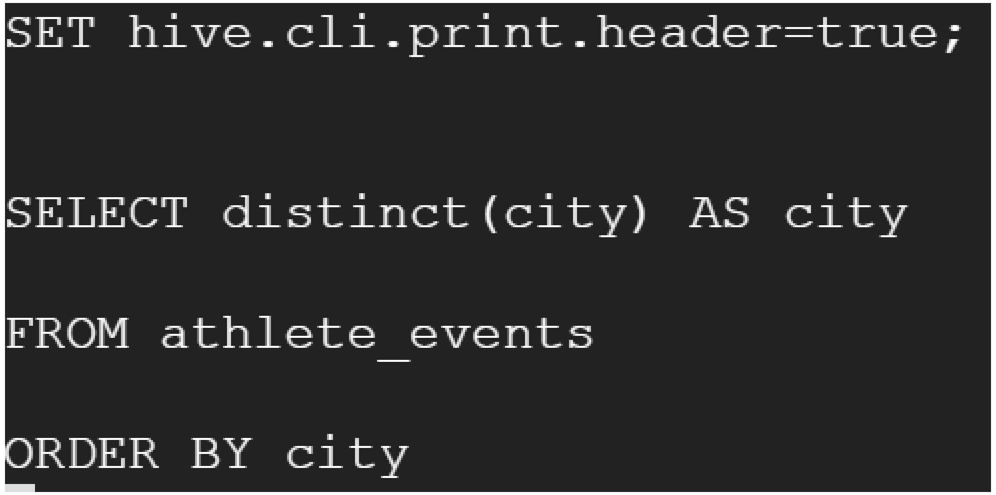


```
cdcnt      mcnt
42         0
```

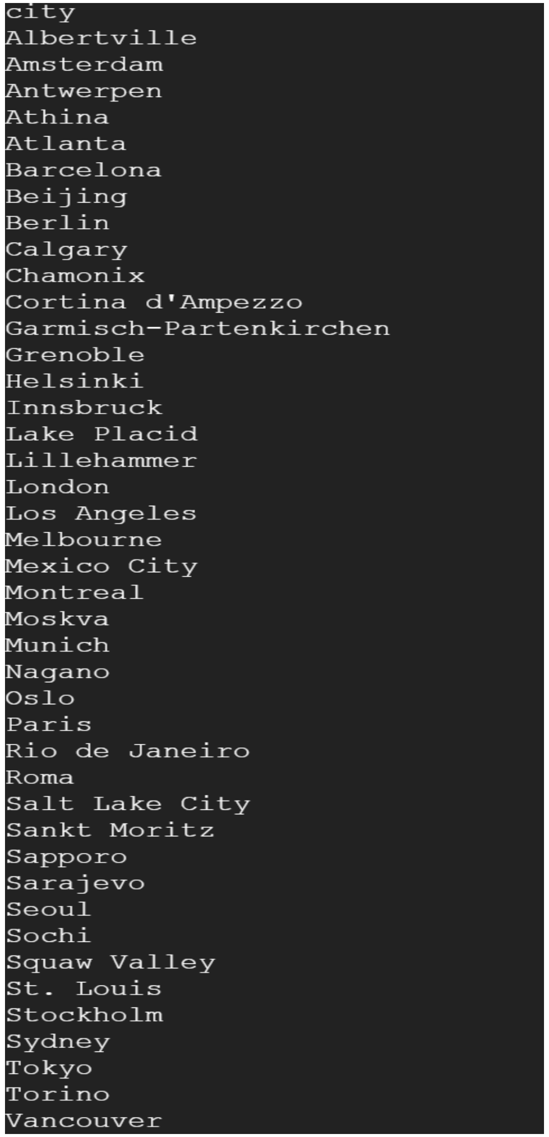


```
SET hive.cli.print.header=true;

SELECT cdcnt, mcnt from

(SELECT count(distinct(city)) cdcnt
FROM athlete_events) a ,

(SELECT count(city) mcnt  from athlete_events
WHERE city is null OR city = 'NA') b
;
```

Column Analytics : City

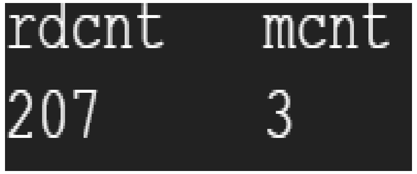


```
SET hive.cli.print.header=true;


SELECT distinct(city) AS city

FROM athlete_events

ORDER BY city
```

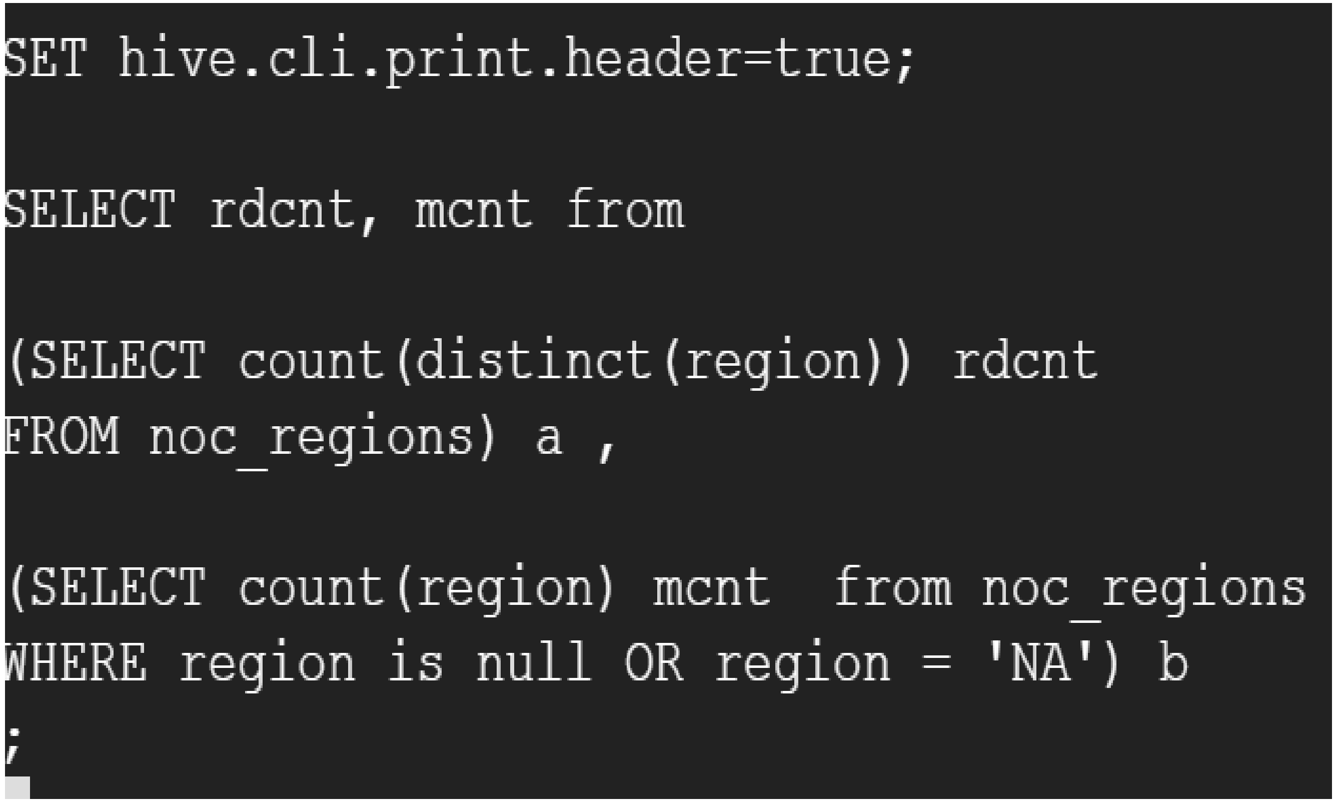


```
city
Albertville
Amsterdam
Antwerpen
Athina
Atlanta
Barcelona
Beijing
Berlin
Calgary
Chamonix
Cortina d'Ampezzo
Garmisch-Partenkirchen
Grenoble
Helsinki
Innsbruck
Lake Placid
Lillehammer
London
Los Angeles
Melbourne
Mexico City
Montreal
Moskva
Munich
Nagano
Oslo
Paris
Rio de Janeiro
Roma
Salt Lake City
Sankt Moritz
Sapporo
Sarajevo
Seoul
Sochi
Squaw Valley
St. Louis
Stockholm
Sydney
Tokyo
Torino
Vancouver
```

Column Analytics : Region

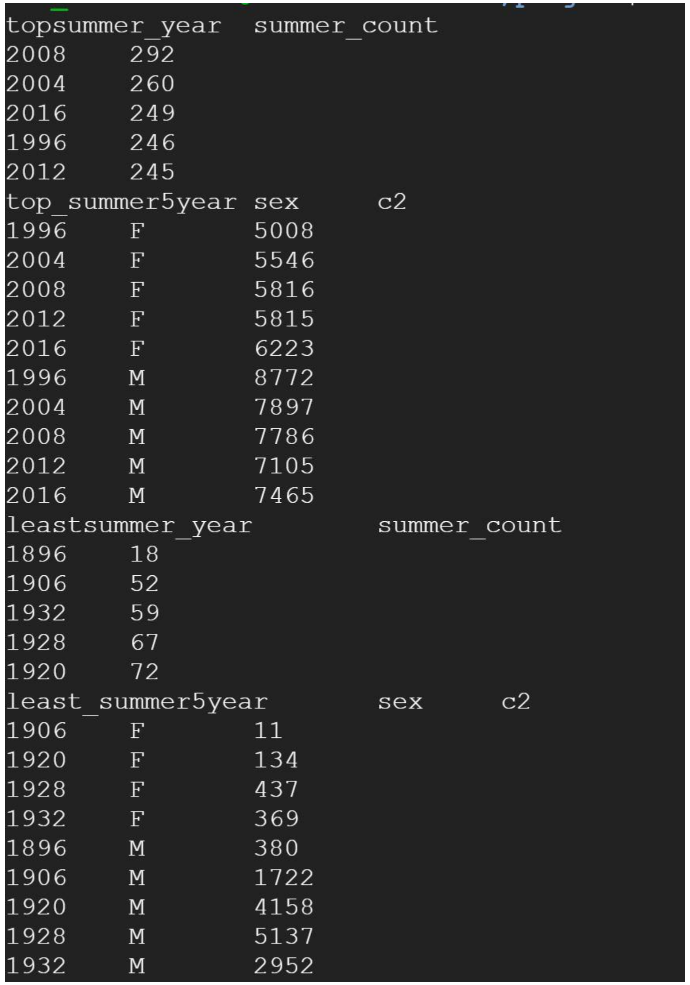


```
rdcnt     mcnt
207       3
```

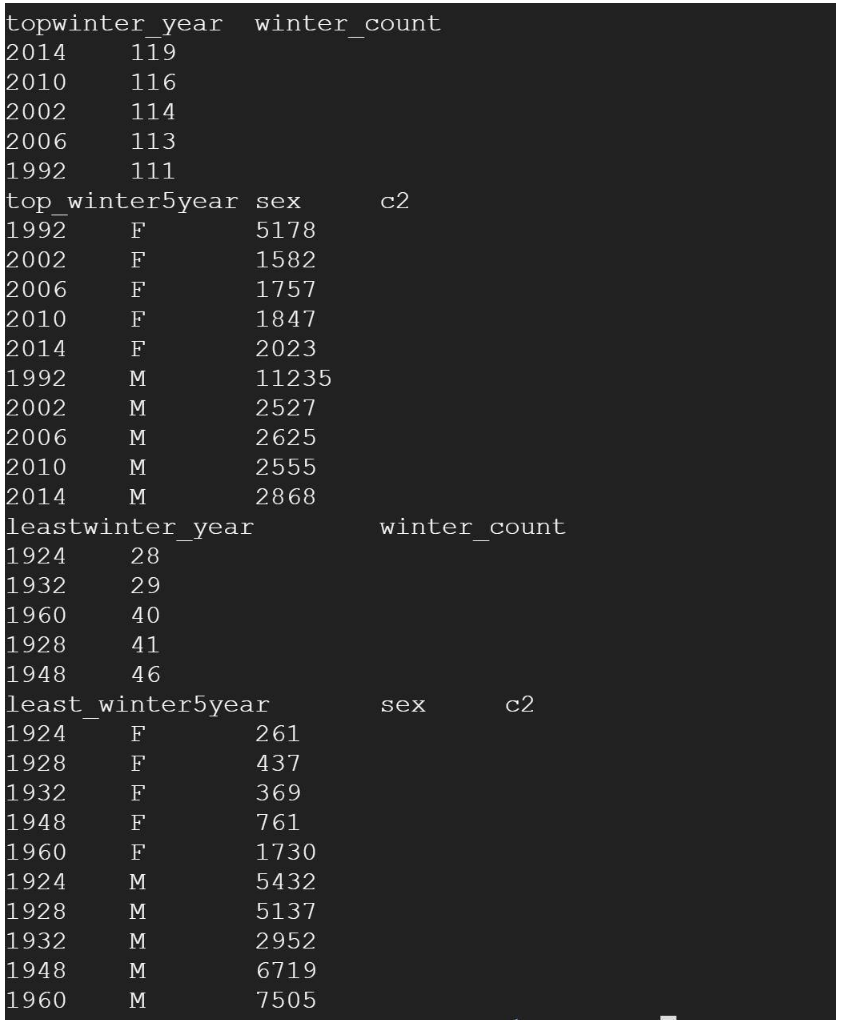


```
SET hive.cli.print.header=true;

SELECT rdcnt, mcnt from

(SELECT count(distinct(region)) rdcnt
FROM noc_regions) a ,

(SELECT count(region) mcnt  from noc_regions
WHERE region is null OR region = 'NA') b
;
```

Top and Bottom 5 Olympics(years) in terms of the number of countries participated and the count of male and female athletes in those years?

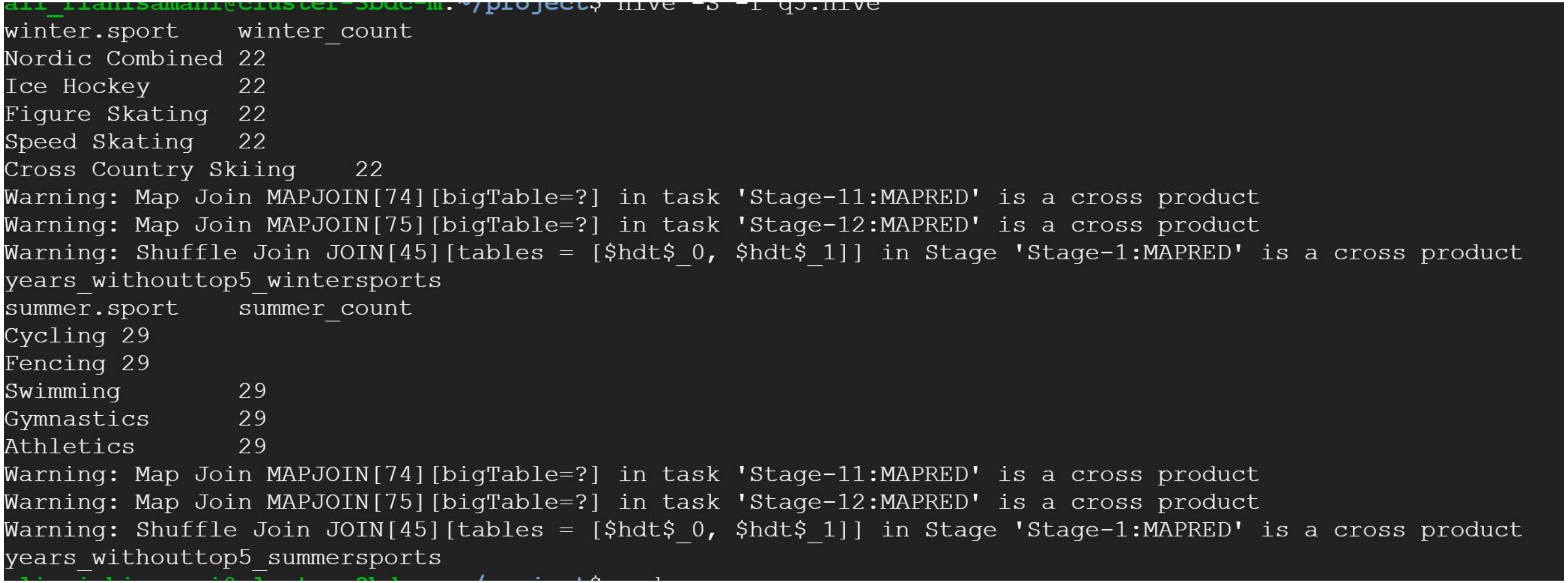


```
topsummer_year  summer_count
2008    292
2004    260
2016    249
1996    246
2012    245
top_summer5year sex     c2
1996    F       5008
2004    F       5546
2008    F       5816
2012    F       5815
2016    F       6223
1996    M       8772
2004    M       7897
2008    M       7786
2012    M       7105
2016    M       7465
leastsummer_year        summer_count
1896    18
1906    52
1932    59
1928    67
1920    72
least_summer5year       sex     c2
1906    F       11
1920    F       134
1928    F       437
1932    F       369
1896    M       380
1906    M       1722
1920    M       4158
1928    M       5137
1932    M       2952
```



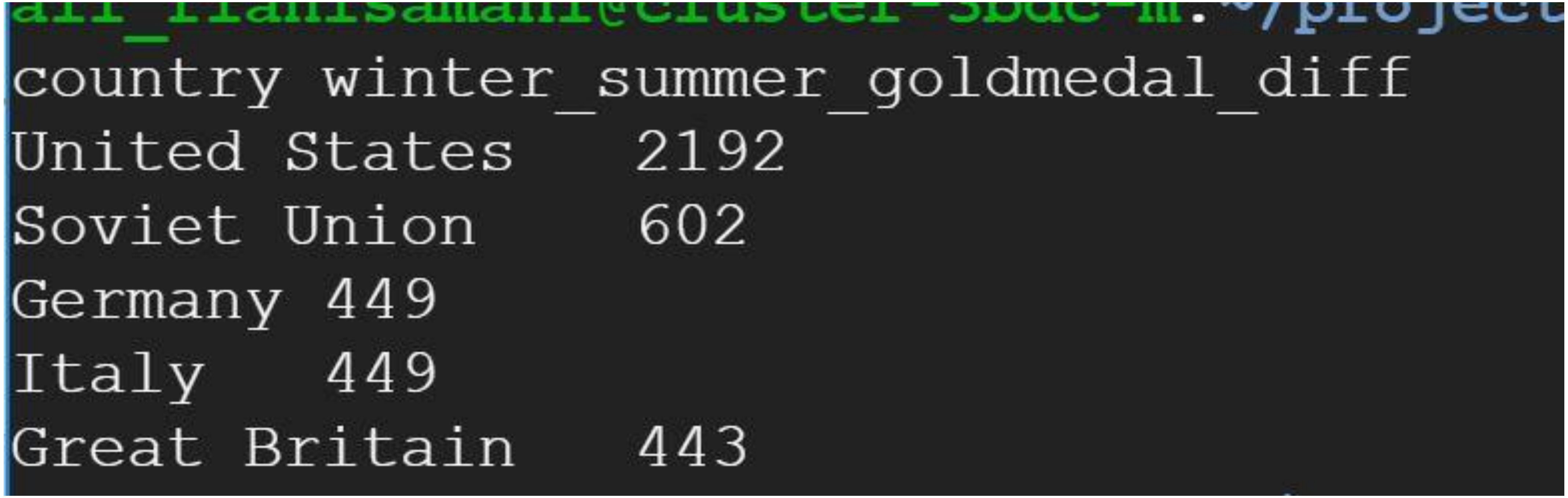
```
topwinter_year  winter_count
2014    119
2010    116
2002    114
2006    113
1992    111
top_winter5year sex     c2
1992    F       5178
2002    F       1582
2006    F       1757
2010    F       1847
2014    F       2023
1992    M       11235
2002    M       2527
2006    M       2625
2010    M       2555
2014    M       2868
leastwinter_year        winter_count
1924    28
1932    29
1960    40
1928    41
1948    46
least_winter5year       sex     c2
1924    F       261
1928    F       437
1932    F       369
1948    F       761
1960    F       1730
1924    M       5432
1928    M       5137
1932    M       2952
1948    M       6719
1960    M       7505
```

Top 5 sports in terms of how many times they were part of an Olympics over the years, and which were the years they were NOT part of an Olympics?



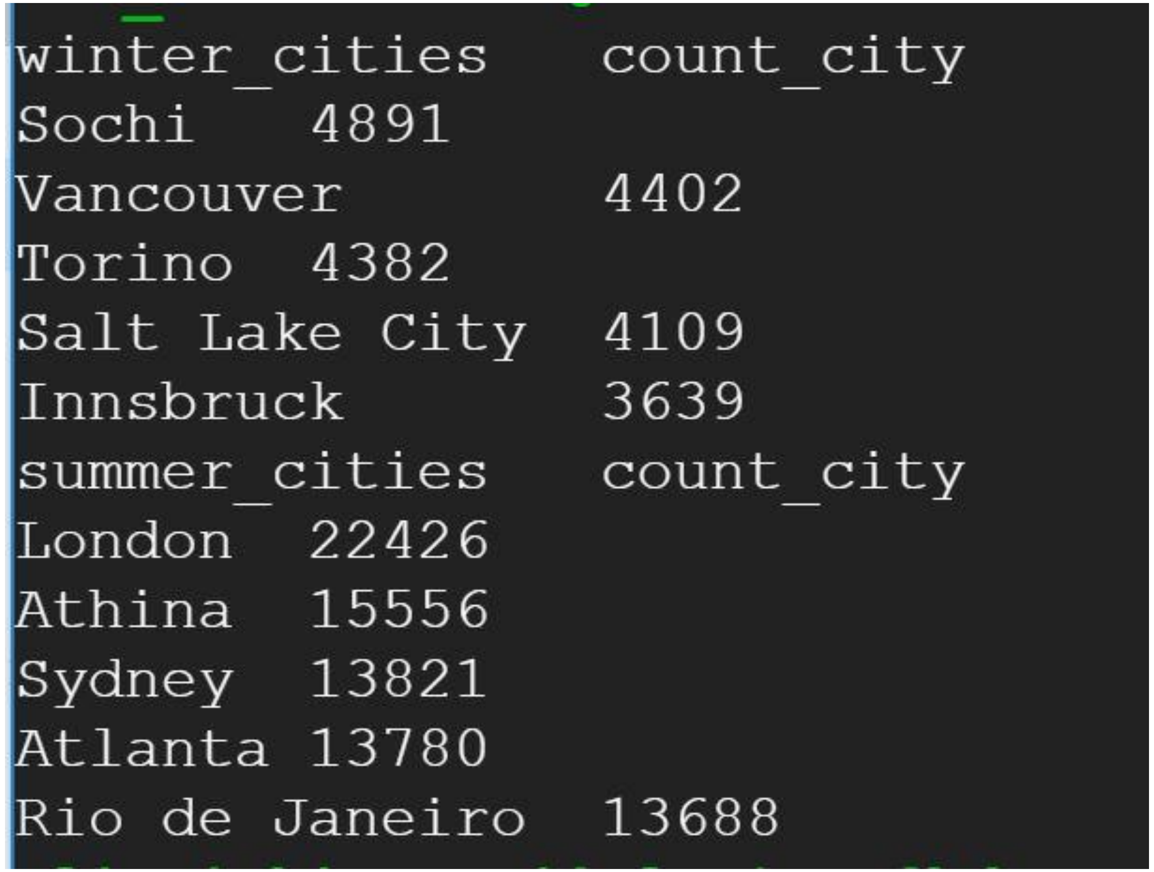
```
ali_rianisamani@cluster-3bdc-m:~/project$ hive -S -f q5.hive
winter.sport    winter_count
Nordic Combined 22
Ice Hockey      22
Figure Skating  22
Speed Skating   22
Cross Country Skiing    22
Warning: Map Join MAPJOIN[74][bigTable=?] in task 'Stage-11:MAPRED' is a cross product
Warning: Map Join MAPJOIN[75][bigTable=?] in task 'Stage-12:MAPRED' is a cross product
Warning: Shuffle Join JOIN[45][tables = [$hdt$_0, $hdt$_1]] in Stage 'Stage-1:MAPRED' is a cross product
years_withouttop5_wintersports
summer.sport    summer_count
Cycling 29
Fencing 29
Swimming        29
Gymnastics      29
Athletics       29
Warning: Map Join MAPJOIN[74][bigTable=?] in task 'Stage-11:MAPRED' is a cross product
Warning: Map Join MAPJOIN[75][bigTable=?] in task 'Stage-12:MAPRED' is a cross product
Warning: Shuffle Join JOIN[45][tables = [$hdt$_0, $hdt$_1]] in Stage 'Stage-1:MAPRED' is a cross product
years_withouttop5_summersports
```

Top 5 countries that had the biggest difference between their summer and winter gold medal counts?



```
ali_fianisamani@cluster-3bdc-m:~/project
country winter_summer_goldmedal_diff
United States   2192
Soviet Union    602
Germany 449
Italy   449
Great Britain   443
```

Top 5 cities which have hosted most number of Olympics in summer and winter respectively?



```
winter_cities   count_city
Sochi   4891
Vancouver       4402
Torino  4382
Salt Lake City  4109
Innsbruck       3639
summer_cities   count_city
London  22426
Athina  15556
Sydney  13821
Atlanta 13780
Rio de Janeiro  13688
```

Data Source and References

- https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

- https://en.wikipedia.org/wiki/Olympic_Games
- https://www.olympic.org/ancient-olympic-games/history-old