```
from pyspark.sql.functions import col
from pyspark.sql.types import IntegerType, DoubleType, BooleanType, DateType
```

```
configs = {"fs.azure.account.auth.type": "OAuth",
"fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenPr
"fs.azure.account.oauth2.client.id": "4bd42c13-9106-406d-835d-4c8cd3ea4a23",
"fs.azure.account.oauth2.client.secret": 'ozs8Q~KFWm6--XH6m2C_nG~NEu6KCQHYXLtZncKQ',
"fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/4c912556-c8b7-4f3f

dbutils.fs.mount(
    source = "abfss://tokyo-olympic-data@tokyoolympicdatakavya.dfs.core.windows.net/",
    mount_point = "/mnt/tokyoolympicdatakavyaupdated",
    extra_configs = configs
)
```

```
Out[33]: True
```

```
%fs
ls "/mnt/tokyoolympicdatakavyaupdated"
```

path	name	size	modificationTime
dbfs:/mnt/tokyoolympicdatakavyaupdated/raw-data/	raw-data/	0	1740227697000
dbfs:/mnt/tokyoolympicdatakavyaupdated/transformed-data/	transformed-data/	0	1740227728000

```
dbutils.fs.ls("/mnt/tokyoolympicdatakavyaupdated/raw-data/")
```

```
Out[46]: [FileInfo(path='dbfs:/mnt/tokyoolympicdatakavyaupdated/raw-data/athletes.csv', nam
 FileInfo(path='dbfs:/mnt/tokyoolympicdatakavyaupdated/raw-data/coaches.csv', name='coaches
 FileInfo(path='dbfs:/mnt/tokyoolympicdatakavyaupdated/raw-data/entriesgender.csv', name='e
 FileInfo(path='dbfs:/mnt/tokyoolympicdatakavyaupdated/raw-data/medals.csv', name='medals.c
 FileInfo(path='dbfs:/mnt/tokyoolympicdatakavyaupdated/raw-data/teams.csv', name='teams.csv
```

```
athletes = spark.read.format("csv").option("header", "true").option("inferSchema","true").load("
coaches = spark.read.format("csv").option("header", "true").option("inferSchema","true").load("/
entriesgenders = spark.read.format("csv").option("header", "true").option("inferSchema","true").
medals = spark.read.format("csv").option("header", "true").option("inferSchema","true").load("/m
teams = spark.read.format("csv").option("header", "true").option("inferSchema","true").load("/mn
```

```
athletes.printSchema()
```

```
root
 |-- PersonName: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- Discipline: string (nullable = true)
```

```
coaches.show()
```

```
+--------------------+--------------------+-----------------+--------+
|                Name|             Country|       Discipline|   Event|
+--------------------+--------------------+-----------------+--------+
|     ABDELMAGID Wael|               Egypt|         Football|    null|
|           ABE Junya|               Japan|       Volleyball|    null|
|       ABE Katsuhiko|               Japan|       Basketball|    null|
|        ADAMA Cherif|       C⍰te d'Ivoire|         Football|    null|
|           AGEBA Yuya|               Japan|       Volleyball|    null|
|AIKMAN Siegfried ...|               Japan|           Hockey|     Men|
|       AL SAADI Kais|             Germany|           Hockey|     Men|
|       ALAMEDA Lonni|              Canada|Baseball/Softball|Softball|
|     ALEKNO Vladimir|Islamic Republic ...|       Volleyball|     Men|
|     ALEKSEEV Alexey|                 ROC|         Handball|   Women|
|ALLER CARBALLO Ma...|               Spain|       Basketball|    null|
|       ALSHEHRI Saad|        Saudi Arabia|         Football|     Men|
|           ALY Kamal|               Egypt|         Football|    null|
| AMAYA GAITAN Fabian|         Puerto Rico|       Basketball|    null|
|    AMO AGUADO Pablo|               Spain|         Football|    null|
|   ANDONOVSKI Vlatko|United States of ...|         Football|   Women|
|        ANNAN Alyson|         Netherlands|           Hockey|   Women|
|  ARNAU CREUS Xavier|               Japan|           Hockey|   Women|
|       ARNOLD Graham|           Australia|         Football|     Men|
|         AXNER Tomas|              Sweden|         Handball|   Women|
+--------------------+--------------------+-----------------+--------+
only showing top 20 rows
```

```
coaches.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- Discipline: string (nullable = true)
 |-- Event: string (nullable = true)
```

```
entriesgenders.show()
```

```
+--------------------+------+----+-----+
|          Discipline|Female|Male|Total|
+--------------------+------+----+-----+
|      3x3 Basketball|    32|  32|   64|
|             Archery|    64|  64|  128|
| Artistic Gymnastics|    98|  98|  196|
|   Artistic Swimming|   105|   0|  105|
|           Athletics|   969|1072| 2041|
|           Badminton|    86|  87|  173|
|   Baseball/Softball|    90| 144|  234|
|          Basketball|   144| 144|  288|
|    Beach Volleyball|    48|  48|   96|
|              Boxing|   102| 187|  289|
|        Canoe Slalom|    41|  41|   82|
|        Canoe Sprint|   123| 126|  249|
|Cycling BMX Frees...|    10|   9|   19|
|  Cycling BMX Racing|    24|  24|   48|
|Cycling Mountain ...|    38|  38|   76|
|        Cycling Road|    70| 131|  201|
|       Cycling Track|    90|  99|  189|
|              Diving|    72|  71|  143|
|          Equestrian|    73| 125|  198|
|             Fencing|   107| 108|  215|
```

```
+--------------------+------+----+-----+
only showing top 20 rows
```

```
entriesgenders.printSchema()
```

```
root
 |-- Discipline: string (nullable = true)
 |-- Female: integer (nullable = true)
 |-- Male: integer (nullable = true)
 |-- Total: integer (nullable = true)
```

```
entriesgenders = entriesgenders.withColumn("Female",col("Female").cast(IntegerType()))\
    .withColumn("Male",col("Male").cast(IntegerType()))\
        .withColumn("Total",col("Total").cast(IntegerType()))
```

```
medals.show()
```

```
+----+--------------------+----+------+------+-----+-------------+
|Rank|         TeamCountry|Gold|Silver|Bronze|Total|Rank by Total|
+----+--------------------+----+------+------+-----+-------------+
|   1|United States of ...|  39|    41|    33|  113|            1|
|   2|People's Republic...|  38|    32|    18|   88|            2|
|   3|               Japan|  27|    14|    17|   58|            5|
|   4|       Great Britain|  22|    21|    22|   65|            4|
|   5|                 ROC|  20|    28|    23|   71|            3|
|   6|           Australia|  17|     7|    22|   46|            6|
|   7|         Netherlands|  10|    12|    14|   36|            9|
|   8|              France|  10|    12|    11|   33|           10|
|   9|             Germany|  10|    11|    16|   37|            8|
|  10|               Italy|  10|    10|    20|   40|            7|
|  11|              Canada|   7|     6|    11|   24|           11|
|  12|              Brazil|   7|     6|     8|   21|           12|
|  13|         New Zealand|   7|     6|     7|   20|           13|
|  14|                Cuba|   7|     3|     5|   15|           18|
|  15|             Hungary|   6|     7|     7|   20|           13|
|  16|   Republic of Korea|   6|     4|    10|   20|           13|
|  17|              Poland|   4|     5|     5|   14|           19|
|  18|      Czech Republic|   4|     4|     3|   11|           23|
|  19|               Kenya|   4|     4|     2|   10|           25|
|  20|              Norway|   4|     2|     2|    8|           29|
+----+--------------------+----+------+------+-----+-------------+
only showing top 20 rows
```

```
medals.printSchema()
```

```
root
 |-- Rank: integer (nullable = true)
 |-- TeamCountry: string (nullable = true)
 |-- Gold: integer (nullable = true)
 |-- Silver: integer (nullable = true)
 |-- Bronze: integer (nullable = true)
 |-- Total: integer (nullable = true)
 |-- Rank by Total: integer (nullable = true)
```

```
#Find top countries with highes number of gold medals

top_gold_medal_countries = medals.orderBy("Gold",ascending = False).select("TeamCountry","Gold'
```

```
+--------------------+----+
|         TeamCountry|Gold|
+--------------------+----+
|United States of ...|  39|
|People's Republic...|  38|
|               Japan|  27|
|       Great Britain|  22|
|                 ROC|  20|
|           Australia|  17|
|         Netherlands|  10|
|              France|  10|
|             Germany|  10|
|               Italy|  10|
|              Canada|   7|
|              Brazil|   7|
|         New Zealand|   7|
|                Cuba|   7|
|             Hungary|   6|
|   Republic of Korea|   6|
|              Poland|   4|
|      Czech Republic|   4|
|               Kenya|   4|
|              Norway|   4|
+--------------------+----+
only showing top 20 rows
```

```
average_entries_by_gender = entriesgenders.withColumn('Avg_Female',
entriesgenders['Female']/entriesgenders['Total']).withColumn('Avg_Male',
entriesgenders['Male']/entriesgenders['Total']
)
average_entries_by_gender.show()
```

```
+--------------------+------+----+-----+--------------------+--------------------+
|          Discipline|Female|Male|Total|         Avg_Female|           Avg_Male|
+--------------------+------+----+-----+--------------------+--------------------+
|      3x3 Basketball|    32|  32|   64|                 0.5|                 0.5|
|             Archery|    64|  64|  128|                 0.5|                 0.5|
| Artistic Gymnastics|    98|  98|  196|                 0.5|                 0.5|
|   Artistic Swimming|   105|   0|  105|                 1.0|                 0.0|
|           Athletics|   969|1072| 2041| 0.4747672709456149| 0.5252327290543851|
|           Badminton|    86|  87|  173|0.49710982658959535| 0.5028901734104047|
|   Baseball/Softball|    90| 144|  234|0.38461538461538464| 0.6153846153846154|
|          Basketball|   144| 144|  288|                 0.5|                 0.5|
|    Beach Volleyball|    48|  48|   96|                 0.5|                 0.5|
|              Boxing|   102| 187|  289|0.35294117647058826| 0.6470588235294118|
|        Canoe Slalom|    41|  41|   82|                 0.5|                 0.5|
|        Canoe Sprint|   123| 126|  249| 0.4939759036144578| 0.5060240963855421|
|Cycling BMX Frees...|    10|   9|   19| 0.5263157894736842|0.47368421052631576|
|  Cycling BMX Racing|    24|  24|   48|                 0.5|                 0.5|
|Cycling Mountain ...|    38|  38|   76|                 0.5|                 0.5|
|        Cycling Road|    70| 131|  201| 0.3482587064676617| 0.6517412935323383|
|       Cycling Track|    90|  99|  189|0.47619047619047616| 0.5238095238095238|
|              Diving|    72|  71|  143| 0.5034965034965035| 0.4965034965034965|
|           Equestrian|    73| 125|  198| 0.3686868686868687| 0.6313131313131313|
|             Fencing|   107| 108|  215|0.49767441860465117| 0.5023255813953489|
+--------------------+------+----+-----+--------------------+--------------------+
```

```
only showing top 20 rows
```

```
athletes.repartition(1).write.mode("overwrite").option("header",'true').csv("dbfs:/mnt/tokyooly
```

Start coding or generate with AI.

```
coaches.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolym
entriesgenders.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/to
medals.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolymp
teams.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolympi
```

```
medals = medals.withColumnRenamed("Rank by Total", "Rank_by_Total")
medals.write.mode("overwrite").parquet("/mnt/tokyoolympicdatakavya/transformed-data/medals/")
```

```
medals.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolymp
```

```
medals.printSchema()
```

```
root
 |-- Rank: integer (nullable = true)
 |-- TeamCountry: string (nullable = true)
 |-- Gold: integer (nullable = true)
 |-- Silver: integer (nullable = true)
 |-- Bronze: integer (nullable = true)
 |-- Total: integer (nullable = true)
 |-- Rank_by_Total: integer (nullable = true)
```