dbutils.fs.mkdirs("/FileStore/tables/football")

[Cmd 2](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/814799057140512)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

dbutils.fs.ls("/FileStore/tables/")

[Cmd 3](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/814799057140508)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

**from** pyspark **import** SparkContext

2

**from** pyspark.sql **import** SparkSession,Window,Row

3

**from** pyspark.sql.functions **import** \*

4

**import** matplotlib.pyplot **as** plt

5

**from** pyspark.sql.types **import** \*

6

Command took 1.79 seconds -- by doaamohamed1497@gmail.com at 2/22/2022, 6:53:37 PM on My Cluster

[Cmd 4](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/814799057140517)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

spark = SparkSession \

2

.builder \

3

.appName("firstSpark") \

4

.getOrCreate()

Command took 0.03 seconds -- by doaamohamed1497@gmail.com at 2/22/2022, 6:55:47 PM on My Cluster

[Cmd 5](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/814799057140511)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

df=spark.read.format("csv").options(header="true").load("dbfs:/FileStore/tables/football/")

2

df.show()

(2) Spark Jobs

+--------+---+------+----------+------------------+--------------+----+----+---+ |Match\_ID|Div|Season| Date| HomeTeam| AwayTeam|FTHG|FTAG|FTR| +--------+---+------+----------+------------------+--------------+----+----+---+ | 1| D2| 2009|2010-04-04| Oberhausen|Kaiserslautern| 2| 1| H| | 2| D2| 2009|2009-11-01| Munich 1860|Kaiserslautern| 0| 1| A| | 3| D2| 2009|2009-10-04| Frankfurt FSV|Kaiserslautern| 1| 1| D| | 4| D2| 2009|2010-02-21| Frankfurt FSV| Karlsruhe| 2| 1| H| | 5| D2| 2009|2009-12-06| Ahlen| Karlsruhe| 1| 3| A| | 6| D2| 2009|2010-04-03| Union Berlin| Karlsruhe| 1| 1| D| | 7| D2| 2009|2009-08-14| Paderborn| Karlsruhe| 2| 0| H| | 8| D2| 2009|2010-03-08| Bielefeld| Karlsruhe| 0| 1| A| | 9| D2| 2009|2009-09-26| Kaiserslautern| Karlsruhe| 2| 0| H| | 10| D2| 2009|2009-11-21| Hansa Rostock| Karlsruhe| 2| 1| H| | 11| D2| 2009|2009-12-19| Greuther Furth| Karlsruhe| 1| 4| A| | 12| D2| 2009|2010-04-16| Koblenz| Karlsruhe| 2| 2| D| | 13| D2| 2009|2010-03-21| Oberhausen| Karlsruhe| 1| 0| H| | 14| D2| 2009|2009-09-14| Cottbus| Karlsruhe| 2| 4| A| | 15| D2| 2009|2010-05-02| Duisburg| Karlsruhe| 0| 1| A| | 16| D2| 2009|2009-10-30|Fortuna Dusseldorf| Karlsruhe| 1| 0| H| | 17| D2| 2009|2010-01-15| Aachen| Karlsruhe| 3| 1| H| | 18| D2| 2009|2009-10-18| Augsburg| Karlsruhe| 1| 1| D| | 19| D2| 2009|2010-02-05| St Pauli| Karlsruhe| 2| 1| H| | 20| D2| 2009|2009-08-24| Munich 1860| Karlsruhe| 1| 3| A| +--------+---+------+----------+------------------+--------------+----+----+---+ only showing top 20 rows

Command took 8.98 seconds -- by doaamohamed1497@gmail.com at 2/22/2022, 6:55:50 PM on My Cluster

[Cmd 6](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/814799057140513)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

old\_cols=df.columns[-3:]

2

new\_cols=["HomeTeamgoals","AwayTeamGoals","FinalResult"]

3

old\_new\_cols=[\*zip(old\_cols,new\_cols)]

4

**for** old\_cols,new\_cols **in** old\_new\_cols :

5

df=df.withColumnRenamed(old\_cols,new\_cols)

6

df.show()

(1) Spark Jobs

+--------+---+------+----------+------------------+--------------+-------------+-------------+-----------+ |Match\_ID|Div|Season| Date| HomeTeam| AwayTeam|HomeTeamgoals|AwayTeamGoals|FinalResult| +--------+---+------+----------+------------------+--------------+-------------+-------------+-----------+ | 1| D2| 2009|2010-04-04| Oberhausen|Kaiserslautern| 2| 1| H| | 2| D2| 2009|2009-11-01| Munich 1860|Kaiserslautern| 0| 1| A| | 3| D2| 2009|2009-10-04| Frankfurt FSV|Kaiserslautern| 1| 1| D| | 4| D2| 2009|2010-02-21| Frankfurt FSV| Karlsruhe| 2| 1| H| | 5| D2| 2009|2009-12-06| Ahlen| Karlsruhe| 1| 3| A| | 6| D2| 2009|2010-04-03| Union Berlin| Karlsruhe| 1| 1| D| | 7| D2| 2009|2009-08-14| Paderborn| Karlsruhe| 2| 0| H| | 8| D2| 2009|2010-03-08| Bielefeld| Karlsruhe| 0| 1| A| | 9| D2| 2009|2009-09-26| Kaiserslautern| Karlsruhe| 2| 0| H| | 10| D2| 2009|2009-11-21| Hansa Rostock| Karlsruhe| 2| 1| H| | 11| D2| 2009|2009-12-19| Greuther Furth| Karlsruhe| 1| 4| A| | 12| D2| 2009|2010-04-16| Koblenz| Karlsruhe| 2| 2| D| | 13| D2| 2009|2010-03-21| Oberhausen| Karlsruhe| 1| 0| H| | 14| D2| 2009|2009-09-14| Cottbus| Karlsruhe| 2| 4| A| | 15| D2| 2009|2010-05-02| Duisburg| Karlsruhe| 0| 1| A| | 16| D2| 2009|2009-10-30|Fortuna Dusseldorf| Karlsruhe| 1| 0| H| | 17| D2| 2009|2010-01-15| Aachen| Karlsruhe| 3| 1| H| | 18| D2| 2009|2009-10-18| Augsburg| Karlsruhe| 1| 1| D| | 19| D2| 2009|2010-02-05| St Pauli| Karlsruhe| 2| 1| H| | 20| D2| 2009|2009-08-24| Munich 1860| Karlsruhe| 1| 3| A| +--------+---+------+----------+------------------+--------------+-------------+-------------+-----------+ only showing top 20 rows

Command took 1.03 seconds -- by doaamohamed1497@gmail.com at 2/22/2022, 6:56:09 PM on My Cluster

[Cmd 7](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/4126770949868974)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)



1

bundesliga= df.filter(   (col("Season") >=2000) & (col("Season") <=2010 ) & (col("Div") =="D1") )

[Cmd 8](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/4126770949868978)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

df\_matches= df\

2

.withColumn("HomeTeamWin" ,when(col("FinalResult") == "H",1).otherwise(0)) \

3

.withColumn("AwayTeamWin", when(col("FinalResult")=="A",1).otherwise(0)) \

4

.withColumn("GameTie",when(col("FinalResult") =="D",1).otherwise(0))

5

df\_matches.show()

[Cmd 9](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/814799057140514)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

2

home=df\_matches.groupby("Season","HomeTeam") .agg(sum("HomeTeamWin").alias("TotalHomeWin"),

3

sum("AwayTeamWin").alias("TotalHomeLoss"),

4

sum("GameTie").alias("TotalHomeTie"),

5

sum("HomeTeamGoals").alias("HomeScoreGoals"),

6

sum("AwayTeamGoals").alias("HomeAgainstGoals") ) \

7

.withColumnRenamed("HomeTeam","Team")

8

home.show()

9

[Cmd 10](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/814799057140515)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

away = df\_matches.groupby('Season', 'AwayTeam') \

2

.agg(sum('AwayTeamWin').alias('TotalAwayWin'),

3

sum('HomeTeamWin').alias('TotalAwayLoss'),

4

sum('GameTie').alias('TotalAwayTie'),

5

sum('AwayTeamGoals').alias('AwayScoredGoals'),

6

sum('HomeTeamGoals').alias('AwayAgainstGoals')) \

7

.withColumnRenamed('AwayTeam', 'Team')

8

away.show()

[Cmd 11](https://community.cloud.databricks.com/?o=1787800780916590#notebook/814799057140506/command/42325345965686)

[Python](https://community.cloud.databricks.com/?o=1787800780916590)

1

window = ['Season']

2

window = Window.partitionBy(window).orderBy(col('WinPct').desc(), col('GoalDifferentials').desc())

3

table = home.join(away, ['Team', 'Season'], 'inner') .withColumn('GoalsScored', col('HomeScoreGoals') + col('HomeAgainstGoals')) \

4

.withColumn('GoalsAgainst', col('HomeAgainstGoals') + col('AwayAgainstGoals')) \

5

.withColumn('GoalDifferentials', col('GoalsScored') - col('GoalsAgainst')) \

6

.withColumn('Win', col('TotalHomeWin') + col('TotalAwayWin')) \

7

.withColumn('Loss', col('TotalHomeLoss') + col('TotalAwayLoss')) \

8

.withColumn('Tie', col('TotalHomeTie') + col('TotalAwayTie')) \

9

.withColumn('WinPct', round((100\* col('Win')/(col('Win') + col('Loss') + col('Tie'))),2)) \

10

.drop('HomeScoredGoals', 'AwayScoredGoals', 'HomeAgainstGoals', 'AwayAgainstGoals') \

11

.drop('TotalHomeWin', 'TotalAwayWin', 'TotalHomeLoss', 'TotalAwayLoss', 'TotalHomeTie', 'TotalAwayTie') \

12

.withColumn('TeamPosition', rank().over(window))

13

14

table\_df = table.filter(col('TeamPosition') == 1).orderBy(asc('Season')).toPandas()

15

table\_df