Spark SQL Project 3: Knowing Each Other

Fields:

```
case class Student (
        name: String,
        age: String,
        gender: String,
        birthCity: String,
        birthCountry: String,
        siblings: String,
        bachelor: String,
        bachelorLocation: String,
        master: String,
        phd: String,
        faoriteSport: String,
        favoriteMusic: String,
        favoriteBook: String,
        favoriteGame: String,
        favoriteFood: String,
        favoritCar: String,
        favoriteMovie: String,
        favoriteActor: String,
        favoriteCity: String,
        favoriteCanadianCity: String,
        yearIT: String,
        language: String,
        hourDailyStudy: String,
        hourDailyRest: String,
        hourDailySports: String,
        hourDailyWorks: String,
        hourDailySleep: String,
        spokenLanguageCount: String,
        favoriteColor: String,
        isWatchPolitics: String, //Boolean,
        isprofessionalSavvy: String, //Boolean,
        isGoodAtSQL : String, //Boolean,
        isGoodAtCloud: String, //Boolean,
        isGoodAtHadoop: String, //Boolean,
        isGoodAtNoSQL: String, //ring, //Boolean,
        isUsingLinkedInLinkedin: String, //Boolean,
```

```
  linkedinContactsCount: String,
  hasJobInterviews: String //Boolean
)
```

Spark Notebook Analytics

Clean data

Understand your data.

1. Boolean fields (Y / N)

- to have only one of the 2 possibilities: Y, N

2. Missing values

- All nuls, missing data to be replaced with None

3 . Clean column Gender

- We should have only 2 values in this column: M or F

4. Constraints for each column

Queries Data - Analytics

Q1: Same city for bachelor and birth location

Q2: No programming language and no interview

Q3: Count number of languages

Q4: Have Master or PhD and job interviews

Q5: Having Master and bachelor and job interviews

Q6: Favorite city to be the same with the born city

Q7: count guys who like cricket but have no favorite movies

Q8: Count all the siblings

Q9: Students with No work experience and no interview

Q10: How many girls have job interview vs guys

(3 cols: girls_jobInterview, guys_jobinterview, substraction_jobInterview)

Q11: Students with No favorite sport but with favorite music and movie

Q12: Top 3 students studying and working most hours

Q13: Top 3 students sleeping and resting most hours

Q14: Order by yearIT, descending and name ascending

Q15 Group students Having the same name length and gender and count them, order by count; for each group find the subtraction of the count between guys and girls

4 cols: guy_same_length_count, girls_same_name_count, length, subtraction_sameLength

Q16 group students with same age, count them and order by count; Also rename count col to count_same_age

Display 2 columns: count (renamed to count_same_age), age

Q17 count All the guys who Like same sport; same for girls

Note: first clean the column you are interested in

Q18: Total of number of languages for girls, same for boys, subtraction between the 2

Q19 Count students knowing same number of language;

Display 2 columns: name ordered ascending and the count renamed to count_languages

Q20 Compare how many hours girls are studying vs guys (percentage)

Q21 Compare how many girls are in the class vs guys (percentage)

Q22 Compare how many hours girls do sports vs boys (percentage)

Q23 Compare how many of hours girls have leisure time vs guys (3 cols: girls_total_leisure, boys_total_leisure, substraction_hours_leisure); leisure time consists in: ourDailyRest, hourDailySports and hourDailySleep

Q24 Compare how many hours girls have no leisure time vs guys

(3 cols:
girls_total_work, boys_total_work, substraction_hours_work)
no leisure time consists in: hourDailyWorks and hourDailyStudy

Q25 Count distinct colors for girls vs distinct color for guys

3 cols: girls_color_count, guys_color_count, substraction_color_count

Q26: Compare how many girls watch politics vs guys

percentage: 2 out of 5 vs 20 out of 30, where

- 2 the females watching politics
- 5 the total females in the class
- 20 the male watching politics
- 30 total males in the class

Q27: How many students like Classical Music

Q28: How many students know SQL and did not use NoSQL and have job interview

Q29: How many students know NoSQL but don’t know SQL

Q30: How many students know cloud but don’t know any programming language

Q31: How many girls know more than 2 languages

Q32: Compare how many girls know what to do in the future vs guys (percentage)

percentage: 2 out of 5 vs 20 out of 30, where

- 2 the females watching politics
- 5 the total females in the class
- 20 the male watching politics
- 30 total males in the class

Q33: How many students do not know any programming language and do not know what to do in the future

Q34: How many students do not know Hadoop but they know SQL

Q35: do not use linkedin and do not know what to do in the future

Q36: choose the favorite place in the world same as the favorite city in canada

Q37: How many students have the favorite city in Canada as Toronto

Q38: How many girls use LinkedIn

Q39: How many students had job interviews in canada and use Linkedin vs How many students had job interviews in canada and don’t use Linkedin

Q40: How many students had job interviews in canada and used either a Programming Language or SQL vs How many students had job interviews in canada and didn’t use neither a Programming Language nor SQL

Q41: How many students had job interviews in canada and used SQL but didn’t use a Programming Language

Q42: How many students had job interviews in canada and used Hadoop

Q43: How many students used Java vs Python

Q44: Describe what are the constraints for each column (e.g. age cannot be < 0 or > 100)

Q45: Each student should come up with one new possible Question

Q46: Who knows most languages boys or guys?

https://stackoverflow.com/questions/43232363/get-min-and-max-from-a-specific-column-scala-spark-dataframe

Q49: Students from same country
Q50: Students from same city

Q51: how many time country with max population in Canada is greater than each of other country population

https://stackoverflow.com/questions/46087420/scala-add-new-column-to-dataframe-by-expression

Code:

```
package model

case class Student (
                    name: String,
                    age: String,
                    gender: String,
                    birthCity: String,
                    birthCountry: String,
                    siblings: String,
                    bachelor: String,
                    bachelorLocation: String,
                    master: String,
                    phd: String,
                    faoriteSport: String,
                    favoriteMusic: String,
                    favoriteBook: String,
                    favoriteGame: String,
                    favoriteFood: String,
                    favoritCar: String,
                    favoriteMovie: String,
                    favoriteActor: String,
                    favoriteCity: String,
                    favoriteCanadianCity: String,
                    yearIT: String,
                    language: String,
                    hourDailyStudy: String,
                    hourDailyRest: String,
```

```
  hourDailySports: String,
  hourDailyWorks: String,
  hourDailySleep: String,
  spokenLanguageCount: String,
  favoriteColor: String,
  isWatchPolitics: String, //Boolean,
  isprofessionalSavvy: String, //Boolean,
  isGoodAtSQL : String, //Boolean,
  isGoodAtCloud: String, //Boolean,
  isGoodAtHadoop: String, //Boolean,
  isGoodAtNoSQL: String, //ring, //Boolean,
  isUsingLinkedInLinkedin: String, //Boolean,
  linkedinContactsCount: String,
  hasJobInterviews: String //Boolean
)
```

```
package exercises

 import model.Student
 import org.apache.spark.sql
 import org.apache.spark.sql.functions._
 import org.apache.spark.sql.types.IntegerType
 import org.apache.spark.sql.{Column, DataFrame, Row, SparkSession}


object StudentsClassAnalytics extends App {
 val spark = SparkSession
   .builder()
   .master("local[*]")
   .appName("Lab 04")
   .getOrCreate()

 val sc = spark.sparkContext
 val sqlContext = spark.sqlContext

 import spark.implicits._


 val dataRDD = sc.textFile("students1.txt").map(l => l.split(",")).map(w =>
Row.fromSeq(w))

 val schema = Seq[Student]().toDF.schema
 val df = spark.createDataFrame(dataRDD, schema)
 val df2 = castColumnsType(df)
```

```
df2.printSchema

//The underscore (must have a space in between function and underscore)
//turns the function into a partially applied function
sqlContext.udf.register("genderColumnClean", genderColumnClean _)
val myExpression = "genderColumnClean(gender)"
val df3 = df2.withColumn("gender", expr(myExpression))
df2.orderBy($"name").show(50)
df3.orderBy($"name").show(50)

val count = df2.count
println (s">>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>> count $count")


//  df2.select($"yearIT").show
//  df2.agg(min($"yearIT"), max($"yearIT")).show
//
//  //count distinct country
//  df2.select($"name", $"birthCountry").show
//  df2.select(countDistinct($"birthCountry").as("distinctCountry")).show
//  //how many students from each country
//  df2.select($"name",
$"birthCountry").groupBy($"birthCountry").agg(count($"name")).show
//  //how many time  country with max population here is grater than each of other
country
//
//
//  val procent ="population/3.0 "
//  val df3 = df2.select($"name",
$"birthCountry").groupBy($"birthCountry").agg(count($"name").as("population").cast(sql.
types.DoubleType).as("population"))
//     //.withColumn("procentPopulation", expr(procent)).show
//
//  df3.show
//  val df4 = df3.withColumn("procent", expr(procent)).show
//
//
//  //Students: +0.5  => devide to max(population) instead of 3
//
//how much leisure times the students have
```

```
 val leisureTime = "hourDailyRest + hourDailySports + hourDailySleep"
 df2.select($"name", $"hourDailyRest", $"hourDailySports",
$"hourDailySleep").withColumn("leisureTime", expr(leisureTime)).show



 /**
   * Cast columns to different type than String; Keep the same name for the column
   * If a value cannot be casted, a null will be returned
   *
   * @param df
   * @return
   */
 def castColumnsType(df: DataFrame) = df.select(
   df.columns.map {
     case age @ "age" => df(age).cast(IntegerType).as(age)
     case siblings @ "siblings" => df(siblings).cast(IntegerType).as(siblings)
     case yearIT @ "yearIT" => df(yearIT).cast(IntegerType).as(yearIT)
     case hourDailyStudy @ "hourDailyStudy" =>
df(hourDailyStudy).cast(IntegerType).as(hourDailyStudy)
     case hourDailyRest @ "hourDailyRest" =>
df(hourDailyRest).cast(IntegerType).as(hourDailyRest)
     case hourDailySports @ "hourDailySports" =>
df(hourDailySports).cast(IntegerType).as(hourDailySports)
     case hourDailyWorks @ "hourDailyWorks" =>
df(hourDailyWorks).cast(IntegerType).as(hourDailyWorks)
     case hourDailySleep @ "hourDailySleep" =>
df(hourDailySleep).cast(IntegerType).as(hourDailySleep)
     case spokenLanguageCount @ "spokenLanguageCount" =>
df(spokenLanguageCount).cast(IntegerType).as(spokenLanguageCount)
     case other         => df(other)
   }: _*
 )

 def booleanColumnClean(df: DataFrame): DataFrame = {
   df
 }

 def genderColumnClean(gender: String): String = gender.trim match {
     case null => null
```

```
    case "MALE" | "Male" | "M" | "m" | "male" => "M"
    case "f" | "female" | "Female" | "F" => "F"
    case _ => null
  }
}
```

```
Tulio,30,M,Sorocaba,Brazil,0,Computer
Science,Sorocaba,None,None,soccer,Rock,1984,Arma3,Pizza,BMW,filme,Keanu
Reeves,Monaco,Toronto,10,Java,2,0,0,0,8,3,green,y,y,y,y,n,y,y,186,n
Amal das,23,M,Kozhikode,India,1,IT,Vadakara,N,N,Cricket,Pop,The power of subconscious
mind,GTA,Biriyani,Tesla,Lights out,Tom cruise,New York,Toronto,1,Java,2,6,1,5,10,3,Blue,N,Y,Y,Y,Y,Y,Y,120,N
Sandeep Singh,24,M,New Delhi,India,1,BTECH,New Delhi,,,Table Tennis,Punjabi,The White Tiger,Sudoku,Afghani
Chicken,Mercedees,The Hobbit,Amir Khan,Dharamshala,Vancouver,2.5,,1,8,1,2,8,3,Turqoise,N,Y,Y,N,N,N,Y,50,N
Smit,21,Male,Junagadh,India,0,B.Tech(IT),Nadiad,,,Cricket,Hollywood,Sherlock Holmes,Counter
Strike,Homemade,Aston Martin,The Pursuit of Happiness,Shah Rukh
Khan,Maldives,Mississauga,1,Andrioid,1,2,1,7,8,3,Black,N,Business,Y,N,N,Y,N,,N
Dhruv,23,M,Amritsar,India,2,Computer
Application,Amritsar,none,none,Weightlifting,Pop,none,Crossfit,Pizza,McLaren,Rocky4,Tom
Cruize,Venice,Vencouver,none,java,2,7,2,8,6,3,black,no,no,yes,no,yes,no,yes,5,no
Deepali,22,Female,Goa,India,1,btech,india,,,basketball,hiphop,Secret,nfs,tacos,porsche,theproposal,ZacEfron,Dubai,
vancouver,2,Java,2,5,0,0,5,3,orange,y,y,y,n,n,n,y,150,n
Vinit,24,M,Nagpur,India,1,Computer Technology,Nagpur,,,Cricket,Bollywood,,,Indian,,,New
York,Toronto,2,SQL,2,2,,4,8,3,Black,N,Y,Y,N,N,N,Y,100,N
Rahul,23,M,Hisar,India,2,IT,Indore,,,swimming,love,,Dota2,Butter Chicken,ToyotaSupra,Interstellar,Leonardo
dicaprio,Hisar,Montreal,0,,,2,0,,6,3,Black,No,No,Yes,No,Yes,No,No,,Yes
Tony,22,M,Pala,India,1,IT,Pala,N,N,Soccer,Pop,Deep
Work,GTA4,Biriyani,BMW,Cars,Mohanlal,Pala,Mississauga,0,Python,2,4,1,8,6,3,Blue,Y,Y,Y,Y,N,N,Y,50,Y
MAHARSHI, 21, MALE,  UDAIPUR, INDIA, 1, IT, UDAIPUR, , ,CRICKET, ROMANTIC, RICH DAD POOR DAD, MINI
MILITIA, TANDORI CHICKEN, TESLA MODEL S, INJUSTICE, TOM CRUISE, PORT BLAIR, VANCOUVER, 0,
JAVA, 3 ,7, 1 ,4 ,7 ,3, BLACK, NO, YES, YES, NO, YES, NO, YES, 10, YES
Achyuth,23,M,hyd,India,2,cs,hyd,,,,,,,,biryani,ferrari,,,Hyd,toronto,2,3,1,10,1,2,6,3,blue,n,y,y,n,y,n,y,50,y
Tejal,21,F,Vadodara,India,3 sisters,Mathematics,Vadodara,,,Vollyball,Rock,Perspectives,Candy Crush,Pizza,Audi,Ek
tha Tiger,Salman Khan,New York,Toronto,0,C++,2,1,0,3,6,3,Red,N,Y,N,N,N,N,N,0,N
Rushabh Raolji, 25, M, Anand, India, 0, IT, Anand, Computer Science, ,Tennis, Pop, The Alchemist, Watch Dogs,
Indian, Massareti, , Jim Carrey, Paris, Toronto, 2, Python, 1, 4, 1, 5, 8, 4, Black, Y, Y, Y, Y, N, Y, Y, 2329, Y
Olha,21,F,Kiev,Ukraine,0,IT,Kiev,,,ping-pong,Indi,Tree Comrades,Hearthstone,Pasta,Mini Cooper,The
Intouchables,Benedict Cumbertbatch,Munich,Toronto,2,X++,4,4,1,6,6,3,dark blue,N,N,Y,Y,Y,N,Y,100,N
Lucicarla,39,F,Olinda,Brazil,2,IT,Recife,,,Swim,Pop,Bible,just dance,BBQ,Volvo,notebook,,NY,Ottawa,17,PL-
SQL,2,2,0,3,6,3,blue,N,Y,Y,N,N,N,Y,50,N
Valéria,38,F,Brasilia,Brazil,3,Administration,Brasilia,,,Soccer,Paradise,Sophie's World,Little Big
Planet,Barbecue,Porsche,Pursuit of Happinnes,Morgan Freeman,Amsterdam,Vancouver,11,unix shell
script,1,1,,3,7,3,Red,Y,Y,Y,Y,Y,Y,Y,227,Y
Jibin,22,M,Riyadh,Saudi Arabia,1,BCA,Bengaluru,,,Football,,Balarama,Football,Puttu,Tesla,Prestige,Hugh
Jackman,Bengaluru,Niagara,1,Malayalam,2,7,2,2,6,5,White,Yes,Yes,Yes,No,No,No,No,,Yes
Johns George,21,M,Kottayam,India,2,IT,Pala,0,0,Cricket,David Guetta, The secret ,GTA 4,Kappa and Beef, BMW,
Fast and Furious, Paul Walker, New York, Toronto Downtown,0,C++,2,8,0,5,8,3,Blue,No,Data
Analysis,Yes,No,No,No,Yes,10,No
```

Shruti, 22, female, Visakhapatnam, India, 1 brother, computer science, Visakhapatnam, big data, , badminton, pop, And the mountains echoed, chess, fried rice, tesla, , shahrukh khan, dubai, Toronto, 6 months, java, 2 hours, 2 hours, 1 hour, 6 hours, 7 hours, 3 languages, blue, yes, yes, yes, no, no, no, yes, 7, not yet
Javier,27,M,Juarez City,Mexico,1,Computer Systems,Mexico,N,N,Soccer,Rock,Soccernomics,Fifa EA Sports,Enchiladas,Toyota,Moneyball,Seth Rogen,Sayulita,Banff,4,Java,2,1,0,0,8,3,Blue,N,Y,Y,N,N,N,Y,286,N
Shweta,21,F,Nawanshahr,India,1,Physics,Amritsar,,,Cricket,Classical,The Fountainhead,Carromboard,Pav Bhaji, Mercedes,Fault in our stars, Akshay Kumar, Sydney, Etobicoke,,Python,1,10,3,7,8,6,Scarlet,y,teaching,n,n,n,n,n,0,y
Shivani,21,F,Amritsar,India,1 brother,Physics,Amritsar,,,Cricket,Rock,Who will cry when you will die,chess, Rajma Rice,Mercedes,Doctor Strange, Benedict, Sydney,Etobicoke,,Python,1,9,2,4,8,4,Pink,yes,instructor,,,,,,y
Supriya Kapoor, 21, female, Patiala, India, 1, , ludhiana, , , badminton, classical, the alchemist, , golgappas, , three idiots, nana patekar, Dubai, Vancover, 0, c++, 2, 6, 0, 2, 7, 3, white, N, Y, Y, N, N, N, N, 0,N
JASPREET,24,M,BHOGPUR,INDIA,2,BSC,JALANDHAR,MSC,,BADMINTON,CLASSICAL,ONE INDIAN GIRL,CLASH OF TITANS,PUNJABI FOOD,BUGATTI,THE ITALIAN JOB,,,MISSISSAUGA,,,,,5,12,5,3,NAVY BLUE,N,Y,N,N,N,N,N,,N
Karan,23,M,Amritsar,India,1,Computer Science,Amritsar,Mathematics,,Badminton,Punjabi,Men of Mathematics,,Punjabi,BMW,Sardar Mohammad,,Amritsar,Toronto,0,C++,1,2,0,,7,3,Red,N,Y,Y,N,N,N,Y,20,N
Ganga,24,F,Delhi,India,1,CSc,Delhi,,,,Sufi,,Candy Crush,Cheese Pasta,Audi,,Amitab Bacchan,Chennai,,2.8,C#,2,1,,2,7,3,White,N,N,Y,Y,N,N,Y,33,N
Chirag,24,M,Anand,India,1,EnTC,Nagpur,,,Cricket,indian,Immoratals of Meluha,CS go,Indian,,,,Nagpur,Toronto,2,C,,,0,,7,3,Black,Yes,No,Yes,No,No,Yes,No,,No
Rizwan,25,M,Bapatla,India,2,Electronics and communications,India,,,Cricket,Melody,hound of the baskervilles,EA Sports,Hyderabd Biryani,Swift,Chak de india,Shahrukh,Hyderabad,Toronto,3,SQL,3,2,2,8,6,4,Green,N,IT,Y,N,N,N,Y,,N
Navdeep kaur,22,f,Nawanshaher,India,1,BTech ECE,mohali,,,volleyball,rock,,candycrush,chinese,,,,mississauga,mississauga,,java,2,2,1,,8,3,royal blue,n,y,n,y,y,n,n,n,n
gurkirat,22,male,amritsar,india,1,IT,amritsar,bigdata,bigdata,badminton,punjabi,bigdata,badminton,italian,rangerover,fastandfurious,paulwalker,toronto,toronto,1,python,6,8,2,8,4,3,black,no,yes,yes,no,yes,no,yes,200,no
Wasif,25,M,Kozhikode,India,2,IT,Trichy,,,Football,Country,Satanic Verses,AngryBirds,Pasta,Porsche,Godfather,Eddie Redmayne,Paris,Toronto,3,C#,2,5,2,4,7,3,Blue,Y,Y,Y,N,N,N,Y,70,Y
Rushil,23,M,Amritsar,India,1,computer science and engineering,India,,,Cricket,Melody,hound of the baskervilles,need for speed,Butter chicken,Mercedes,Interstellar,leonardo dicaprio,Amritsar,Toronto,1,Java,2,2,2,8,7,3,Navy blue,N,IT,Y,Y,N,N,Y,7,N
Rafael,33,Male,Sao Bernardo,Brazil,1,Administration,Sao Paulo,None,None,Snowboarding,Country Music,Bible,Monopoly,Sushi,Mercedez,Vanilla Sky,Robert Downie Jr,Ilha Bela,Toronto,17,3,4,1,1,0,7,3,White,Y,N,N,Y,Y,N,Y,1242,Y
Siva,25,M,Ongole,India,1,Mechanical,Bapatla,bigdata,None,cricket,classical,1984,counterstrike,biryani,Ferrari,3idiots,amirkhan,greece,Toronto,3,csharp,2,2,1,8,8,3,blue,y,y,y,y,n,y,y,186,y
Tejal,25,f,Pune,india,0,IT,pune,,,,badminton,rock,candycrush,Tandoori,BMW,Insideout,Prabhas,Hawaii,Mississauga,2,Java,1,8,1,2,6,7,Black,n,y,y,y,n,n,y,50,n
Ram,25,m,suryapet,india,1,CSE,bhongir,CSE,,football,melody,arcade,Indian,mini,mohabbatein,Hrithik,Pune,vancouver,2,Java,5,6,0,0,6,3,Blue,y,y,y,y,n,n,y,,25,n
Razikh,25,M,Hyderabad,India,1,Electrical and Electronics,India,,,Cricket,upbeat and happy songs,Hit Refresh – Satya Nadella,Starcraft,Tandoori Chicken,Range Rover Evoque,Secret Superstar ,Aamir Khan,Hyderabad,Toronto,3,python,0.5,5,2,3,8,4,White,Y,Y,Y,Y,Y,N,Y,9980,N