EXAMPLE 1 - Creating Dataset from case class

```
//import $exclude.`org.slf4j:slf4j-log4j12`, $ivy.`org.slf4j:slf4j-nop:1.7.21`, $ivy.`org.slf4j:log4j-over-slf4j:1.7.21` // for cleaner logs
//import $ivy.`org.apache.spark::spark-sql:2.0.2` // adjust spark version - spark >= 1.6 should be fine, possibly >= 1.3 too
//import $ivy.`org.jupyter-scala::spark:0.4.0-RC1` // allows to create SparkContext-s aware of the jupyter-scala kernel
//import jupyter.spark._
//import sqlContext.implicits._
import org.apache.spark.sql.SQLContext
import org.apache.spark.SparkContext


//sparkInit()
val sparkSession = SparkSession.builder.
  appName("SBTB").
  master("local[*]").getOrCreate()
```

Took: 742 milliseconds, at 2017-6-14 16:24

```
// EXAMPLE 1

import org.apache.spark.sql.{Dataset, DataFrame, SparkSession}
import org.apache.spark.sql.functions._
import org.apache.spark.sql._
```

Took: 716 milliseconds, at 2017-6-14 16:24

```
val df = Seq(("Yoda",             "Obi-Wan Kenobi"),
             ("Anakin Skywalker", "Sheev Palpatine"),
             ("Luke Skywalker",   "Han Solo, Leia Skywalker"),
             ("Leia Skywalker",   "Obi-Wan Kenobi"),
             ("Sheev Palpatine",  "Anakin Skywalker"),
             ("Han Solo",         "Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"),
             ("Obi-Wan Kenobi",   "Yoda, Qui-Gon Jinn"),
             ("R2-D2",            "C-3PO"),
             ("C-3PO",            "R2-D2"),
             ("Darth Maul",       "Sheev Palpatine"),
             ("Chewbacca",        "Han Solo"),
             ("Lando Calrissian", "Han Solo"),
             ("Jabba",            "Boba Fett")
            )
.toDF("name", "friends")
```

Took: 1 second 893 milliseconds, at 2017-6-14 16:24

```
case class Friends(name: String, friends: String)
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 720 milliseconds, at 2017-6-14 16:24

```
val friends_ds = df.as[Friends]
```

Took: 891 milliseconds, at 2017-6-14 16:24

```
friends_ds.show()
```

Took: 884 milliseconds, at 2017-6-14 16:24

```
case class Friends_Missing(Who: String, friends: Option[String])
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 554 milliseconds, at 2017-6-14 16:24

```
val ds_missing = Seq(
  ("Yoda",             Some("Obi-Wan Kenobi")),
  ("Anakin Skywalker", Some("Sheev Palpatine")),
  ("Luke Skywalker",   None),
  ("Leia Skywalker",   Some("Obi-Wan Kenobi")),
  ("Sheev Palpatine",  Some("Anakin Skywalker")),
  ("Han Solo",         Some("Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi")))
.toDF("Who", "friends")
.as[Friends_Missing]
```

Took: 730 milliseconds, at 2017-6-14 16:24

```
ds_missing.show()
```

Took: 702 milliseconds, at 2017-6-14 16:24

EXAMPLE 2 - Reading data from CSV and converting it into a Dataset

```
// EXAMPLE 2 - Reading data from CSV and converting it into a Dataset


// First, we define a case class called "Characters"
case class Characters(name: String,
                      height: Double,
                      weight: Option[String],
                      eyecolor: Option[String],
                      haircolor: Option[String],
                      jedi: String,
                      species: String)
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 493 milliseconds, at 2017-6-14 16:24

```
val characters_ds: Dataset[Characters] = sparkSession
  .read
  .option("header", "true")
  .option("delimiter", ",")
  .option("inferSchema", "true")
  .csv("starwars1.csv")
  .as[Characters]
```

Took: 1 second 740 milliseconds, at 2017-6-14 16:24

```
characters_ds.select("name", "weight").where("weight != 'NA'").where("weight > 100").toDF("name", "terribly fat").show()
```

Took: 1 second 128 milliseconds, at 2017-6-14 16:24

```
characters_ds.select("name","weight").where("weight != 'NA'").where("weight < 60").toDF("name","terribly skinny").show()
```

Took: 738 milliseconds, at 2017-6-14 16:24

```
characters_ds.show()
```

Took: 718 milliseconds, at 2017-6-14 16:24

EXAMPLE 3 - Non-expected missing values in data

```
case class Characters_BadType(name: String,
                      height: String,
                      weight: String,
                      eyecolor: String,
                      haircolor: String,
                      jedi: String,
                      species: String)
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 499 milliseconds, at 2017-6-14 16:24

```
val characters_BadType_ds: Dataset[Characters_BadType] = sparkSession
  .read
  .option("header", "true")
  .option("delimiter", ",")
  .option("inferSchema", "true")
  .csv("starwars1.csv")
  .as[Characters_BadType]
```

Took: 603 milliseconds, at 2017-6-14 16:24

```
characters_BadType_ds.show()
```

Took: 630 milliseconds, at 2017-6-14 16:24

```
val characters_BadType_ds2 = characters_BadType_ds.filter(x=> x.jedi=="no_jedi")
```

Took: 488 milliseconds, at 2017-6-14 16:24

```
characters_BadType_ds2.show()
```

Took: 665 milliseconds, at 2017-6-14 16:24

```
characters_BadType_ds2.filter(x=> x.haircolor=="brown").show()
```

Took: 654 milliseconds, at 2017-6-14 16:24

```
characters_BadType_ds2.filter(x=> x.weight !="NA").filter(x=> x.weight.toInt>79).show()
```

Took: 751 milliseconds, at 2017-6-14 16:24

```
characters_BadType_ds2.filter(x=> x.weight!="NA" && x.weight.toInt>79).show()
```

Took: 603 milliseconds, at 2017-6-14 16:24

EXAMPLE 4 - Defining the schema of the DataFrame manually

```
import org.apache.spark.sql.types.{StructType, StructField, StringType, IntegerType, DoubleType}
```

Took: 366 milliseconds, at 2017-6-14 16:24

```
val DF_schema = StructType(Array(
  StructField("name", StringType, false),
  StructField("gender", StringType, false),
  StructField("height", DoubleType, false),
  StructField("weight", StringType, false),
  StructField("eyecolor", StringType, false),
  StructField("haircolor", StringType, false),
  StructField("jedi", StringType, false),
  StructField("species", StringType, false)))
```

Took: 519 milliseconds, at 2017-6-14 16:24

```
DF_schema.printTreeString
```

```
val characters1_df = sparkSession
  .read
  .format("com.databricks.spark.csv")
  .option("header", "true")
  .option("delimiter", ",")
  .schema(DF_schema)
  .csv("starwars1.csv")
```

Took: 422 milliseconds, at 2017-6-14 16:24

```
characters1_df.show()
```

Took: 594 milliseconds, at 2017-6-14 16:24

```
characters1_df.printSchema
```

Took: 542 milliseconds, at 2017-6-14 16:24

```
characters1_df.filter($"weight"<75).show()
```

Took: 691 milliseconds, at 2017-6-14 16:24

Joining Datasets

Example 1 - Inner Join

```
val bad_join_df = characters_ds.join(friends_ds, characters_ds.col("name") === friends_ds.col("name"))
```

Took: 484 milliseconds, at 2017-6-14 16:24

```
bad_join_df.show()
```

Took: 872 milliseconds, at 2017-6-14 16:24

```
bad_join_df.select($"name")
```

```
val sw_df = characters_ds.join(friends_ds, Seq("name"))
```

Took: 449 milliseconds, at 2017-6-14 16:24

```
sw_df.show()
```

Took: 669 milliseconds, at 2017-6-14 16:24

```
case class SW(name: String,
              height: Double,
              weight: Option[String],
              eyecolor: Option[String],
              haircolor: Option[String],
              jedi: String,
              species: String,
              friends: String)
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 383 milliseconds, at 2017-6-14 16:24

```
val sw_ds = sw_df.as[SW]
```

Took: 521 milliseconds, at 2017-6-14 16:24

```
sw_ds.show()
```

Took: 632 milliseconds, at 2017-6-14 16:24

Example 2 - Other Joins

```
characters_ds.join(ds_missing, characters_ds.col("name") === ds_missing.col("Who"), "left_outer").show()
```

Took: 579 milliseconds, at 2017-6-14 16:24

Selecting Columns

```
sw_ds.map(x => (x.name, x.weight))
```

1

Took: 1 second 149 milliseconds, at 2017-6-14 16:24

_1	_2
"Anakin Skywalker"	"84"
"Luke Skywalker"	"77"
"Leia Skywalker"	"49"
"Obi-Wan Kenobi"	"77"
"Han Solo"	"80"
"Sheev Palpatine"	"75"
"R2-D2"	"32"
"C-3PO"	"75"
"Yoda"	"17"
"Darth Maul"	"80"
"Chewbacca"	"112"
"Jabba"	"NA"
"Lando Calrissian"	"79"

```
sw_ds.map(x => (x.name, x.weight)).show()
```

Took: 896 milliseconds, at 2017-6-14 16:24

```
sw_ds.select("name", "weight")
```

1

Took: 744 milliseconds, at 2017-6-14 16:24

name	weight
"Anakin Skywalker"	"84"
"Luke Skywalker"	"77"
"Leia Skywalker"	"49"
"Obi-Wan Kenobi"	"77"
"Han Solo"	"80"
"Sheev Palpatine"	"75"
"R2-D2"	"32"
"C-3PO"	"75"
"Yoda"	"17"
"Darth Maul"	"80"
"Chewbacca"	"112"
"Jabba"	"NA"
"Lando Calrissian"	"79"

```
sw_ds.select($"name".as[String], $"weight".as[String])
```

1

Took: 796 milliseconds, at 2017-6-14 16:24

name	weight
"Anakin Skywalker"	"84"
"Luke Skywalker"	"77"
"Leia Skywalker"	"49"
"Obi-Wan Kenobi"	"77"
"Han Solo"	"80"
"Sheev Palpatine"	"75"
"R2-D2"	"32"
"C-3PO"	"75"
"Yoda"	"17"
"Darth Maul"	"80"
"Chewbacca"	"112"
"Jabba"	"NA"
"Lando Calrissian"	"79"

```
sw_ds.select($"name".as[String], $"weight".as[String]).show()
```

Took: 648 milliseconds, at 2017-6-14 16:24

```
case class NameWeight(name: String, weight: String)
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 447 milliseconds, at 2017-6-14 16:24

```
sw_ds.map(x => NameWeight(x.name, x.weight))
```

```
sw_ds.map(x => NameWeight(x.name, x.weight)).show()
```

```
sw_ds.select("name", "weight").as[NameWeight]
```

1

Took: 766 milliseconds, at 2017-6-14 16:24

name	weight
"Anakin Skywalker"	"84"
"Luke Skywalker"	"77"
"Leia Skywalker"	"49"
"Obi-Wan Kenobi"	"77"
"Han Solo"	"80"
"Sheev Palpatine"	"75"
"R2-D2"	"32"
"C-3PO"	"75"
"Yoda"	"17"
"Darth Maul"	"80"
"Chewbacca"	"112"
"Jabba"	"NA"
"Lando Calrissian"	"79"

```
sw_ds.select("name", "weight").as[NameWeight].show()
```

Took: 669 milliseconds, at 2017-6-14 16:24

```
sw_ds.select($"name".as[String], $"weight".as[String]).as[NameWeight]
```

1

Took: 641 milliseconds, at 2017-6-14 16:24

name	weight
"Anakin Skywalker"	"84"
"Luke Skywalker"	"77"
"Leia Skywalker"	"49"
"Obi-Wan Kenobi"	"77"
"Han Solo"	"80"
"Sheev Palpatine"	"75"
"R2-D2"	"32"
"C-3PO"	"75"
"Yoda"	"17"
"Darth Maul"	"80"
"Chewbacca"	"112"
"Jabba"	"NA"
"Lando Calrissian"	"79"

```
sw_ds.select($"name".as[String], $"weight".as[String]).as[NameWeight].show()
```

Took: 702 milliseconds, at 2017-6-14 16:24

Renaming Columns

```
sw_ds.withColumnRenamed("name", "Name")
```

1

Name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"

Took: 557 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumnRenamed("name", "Name").show()
```

Took: 624 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumnRenamed("name", "Who").withColumnRenamed("jedi", "Religion").show()
```

Took: 575 milliseconds, at 2017-6-14 16:24

```
sw_ds.toDF(Seq("Name", "Gender", "Height", "Weight", "Eyecolor", "Haircolor", "Skincolor", "Homeland", "Born", "Died", "Jedi", "Species", "Weapon", "Friends"): _*).as[SW]
```

1

Name	Gender	Height	Weight	Eyecolor	Haircolor	Skincolor	Homeland	Born	Died	Jedi	Species	Weapon	Friends
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"

Took: 750 milliseconds, at 2017-6-14 16:24

```
sw_ds.toDF(Seq("Name", "Gender", "Height", "Weight", "Eyecolor", "Haircolor", "Skincolor", "Homeland", "Born", "Died", "Jedi", "Species", "Weapon", "Friends"): _*).as[SW].show()
```

Took: 677 milliseconds, at 2017-6-14 16:24

```
case class SW2(WHO: String,
               height: Double,
               weight: Option[String],
               eyecolor: Option[String],
               haircolor: Option[String],
               skincolor: String,
               homeland: String,
               born: String,
               died: String,
               jedi: String,
               species: String,
               weapon: String,
               friends: String)
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 380 milliseconds, at 2017-6-14 16:24

```
sw_ds.toDF(Seq("WHO", "Gender", "Height", "Weight", "Eyecolor", "Haircolor", "Skincolor", "Homeland", "Born", "Died", "Jedi", "Species", "Weapon", "Friends"): _*).as[SW2]
```

1

WHO	Gender	Height	Weight	Eyecolor	Haircolor	Skincolor	Homeland	Born	Died	Jedi	Species	Weapon	Friends
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"

Took: 758 milliseconds, at 2017-6-14 16:24

Adding New Columns

```
sw_ds.withColumn("count", lit(1))
```

1

name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends	count
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"	1
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"	1
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"	1
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"	1
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"	1
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"	1
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"	1
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"	1
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"	1
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"	1
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"	1
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"	1
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"	1

Took: 734 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("count", lit(1)).show()
```

Took: 696 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("log_weight", log("weight"))
```

1

name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends	log_weight
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"	4.430816798843313
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"	4.343805421853684
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"	3.8918202981106265
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"	4.343805421853684
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"	4.382026634673881
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"	4.31748811353631
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"	3.4657359027997265
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"	4.31748811353631
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"	2.833213344056216
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"	4.382026634673881
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"	4.718498871295094
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"	
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"	4.3694478524670215

Took: 720 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("log_weight", log("weight")).show()
```

Took: 674 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("BMI", sw_ds("weight")/(sw_ds("height")*sw_ds("height")/10000))
```

1

name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends	BMI
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"	237664.10140334995
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"	260275.82477014605
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"	217777.77777777778
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"	232459.84784446322
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"	246913.58024691357
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"	250593.07026629688
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"	347222.22222222225
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"	268923.2313815483
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"	390266.2993572084
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"	261224.48979591837
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"	215450.90797168363
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"	
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"	249337.2048983714

Took: 647 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("BMI", sw_ds("weight")/(sw_ds("height")*sw_ds("height")/10000)).show()
```

Took: 615 milliseconds, at 2017-6-14 16:24

User-defined function

```
import scala.reflect.runtime.universe.TypeTag
def createTuple2[Type_x: TypeTag, Type_y: TypeTag] = udf[(Type_x, Type_y), Type_x, Type_y]((x: Type_x, y: Type_y) => (x, y))
```

Took: 794 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("Jedi_Species", createTuple2[String, String].apply(sw_ds("jedi"), sw_ds("species")))
```

1

name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends	Jedi_Species
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"	{"_1":"jedi","_2":"human"}
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"	{"_1":"jedi","_2":"human"}
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"	{"_1":"no_jedi","_2":"human"}
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"	{"_1":"jedi","_2":"human"}
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"	{"_1":"no_jedi","_2":"human"}
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"	{"_1":"no_jedi","_2":"human"}
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"	{"_1":"no_jedi","_2":"droid"}
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"	{"_1":"no_jedi","_2":"droid"}
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"	{"_1":"jedi","_2":"yoda"}
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"	{"_1":"no_jedi","_2":"dathomirian"}
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"	{"_1":"no_jedi","_2":"wookiee"}
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"	{"_1":"no_jedi","_2":"hutt"}
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"	{"_1":"no_jedi","_2":"human"}

Took: 807 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("Jedi_Species", createTuple2[String, String].apply(sw_ds("jedi"), sw_ds("species"))).show()
```

Took: 835 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("Name_Weight", createTuple2[String, Option[Integer]].apply(sw_ds("name"), sw_ds("weight")))
```

1

name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends	Name_Weight
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"	{"_1":"Anakin Skywalker","_2":84}
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"	{"_1":"Luke Skywalker","_2":77}
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"	{"_1":"Leia Skywalker","_2":49}
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"	{"_1":"Obi-Wan Kenobi","_2":77}
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"	{"_1":"Han Solo","_2":80}
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"	{"_1":"Sheev Palpatine","_2":75}
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"	{"_1":"R2-D2","_2":32}
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"	{"_1":"C-3PO","_2":75}
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"	{"_1":"Yoda","_2":17}
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"	{"_1":"Darth Maul","_2":80}
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"	{"_1":"Chewbacca","_2":112}
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"	{"_1":"Jabba"}
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"	{"_1":"Lando Calrissian","_2":79}

Took: 714 milliseconds, at 2017-6-14 16:24

```
sw_ds.withColumn("Name_Weight", createTuple2[String, Option[Integer]].apply(sw_ds("name"), sw_ds("weight"))).show()
```

Took: 684 milliseconds, at 2017-6-14 16:24

Filtering Rows

```
sw_ds.filter(x => x.species=="human")
```

1

name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"

Took: 668 milliseconds, at 2017-6-14 16:24

```
sw_ds.filter(x => x.species=="human").show()
```

Took: 721 milliseconds, at 2017-6-14 16:24

```
sw_ds.filter(x => x.eyecolor== "brown").show()
```

Took: 625 milliseconds, at 2017-6-14 16:24

```
sw_ds.filter(x => x.eyecolor == Some("brown")).show()
```

Took: 730 milliseconds, at 2017-6-14 16:24

```
sw_ds.filter(x => x.eyecolor.getOrElse("") == "brown").show()
```

Took: 586 milliseconds, at 2017-6-14 16:24

```
sw_ds.filter(x =>  x.height<100).show()
```

Took: 826 milliseconds, at 2017-6-14 16:24

```
sw_ds.filter(x => x.weight match {case Some(y) => y.toInt>=79
                                  case None => false} ).filter("weight != 'NA'").show()
```

Took: 723 milliseconds, at 2017-6-14 16:24

GroupBy and Aggregating

```
sw_ds.groupByKey(_.species).agg(max($"height").as[Double], min($"height").as[Double], mean($"weight").as[Double], count($"species").as[Long] )
```

1

Took: 2 seconds 248 milliseconds, at 2017-6-14 16:25

value	max(height)	min(height)	avg(weight)	count(species)
"hutt"	3.9	3.9		1
"human"	1.88	1.5	74.42857142857143	7
"dathomirian"	1.75	1.75	80	1
"yoda"	0.66	0.66	17	1
"wookiee"	2.28	2.28	112	1
"droid"	1.67	0.96	53.5	2

```
sw_ds.groupByKey(_.species).agg(max($"height").as[Double], min($"height").as[Double], mean($"weight").as[Double], count($"species").as[Long] ).show()
```

Took: 927 milliseconds, at 2017-6-14 16:26

```
sw_ds.groupByKey(_.eyecolor).agg(mean($"weight").as[Double])
```

1

Took: 1 second 461 milliseconds, at 2017-6-14 16:26

key	avg(weight)
{"value":"yellow"}	80
{"value":"NA"}	53.5
{"value":"bluegray"}	77
{"value":"brown"}	56.25
{"value":"blue"}	87

```
sw_ds.groupByKey(_.eyecolor).agg(mean($"weight").as[Double]).show()
```

Took: 908 milliseconds, at 2017-6-14 16:26

GroupBy multiple keys

```
sw_ds.groupByKey(x=>(x.species, x.jedi, x.haircolor)).agg(mean($"weight").as[Double], count($"species").as[Long])
```

1

Took: 1 second 340 milliseconds, at 2017-6-14 16:27

key	avg(weight)	count(species)
{"_1":"hutt","_2":"no_jedi","_3":"none"}		1
{"_1":"human","_2":"no_jedi","_3":"blank"}	79	1
{"_1":"dathomirian","_2":"no_jedi","_3":"none"}	80	1
{"_1":"human","_2":"no_jedi","_3":"red"}	75	1
{"_1":"wookiee","_2":"no_jedi","_3":"brown"}	112	1
{"_1":"human","_2":"no_jedi","_3":"brown"}	64.5	2
{"_1":"yoda","_2":"jedi","_3":"brown"}	17	1
{"_1":"human","_2":"jedi","_3":"auburn"}	77	1
{"_1":"droid","_2":"no_jedi","_3":"NA"}	53.5	2
{"_1":"human","_2":"jedi","_3":"blond"}	80.5	2

```
sw_ds.groupByKey(x=>(x.species, x.jedi, x.haircolor)).agg(mean($"weight").as[Double], count($"species").as[Long] ).show()
```

Took: 973 milliseconds, at 2017-6-14 16:28

Sorting by rows

```
sw_ds.orderBy($"species".desc, $"weight")
```

1

name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"

Took: 940 milliseconds, at 2017-6-14 16:29

```
sw_ds.orderBy($"species".desc, $"weight").show()
```

Took: 766 milliseconds, at 2017-6-14 16:29

Appending Datasets

```
sw_ds.union(sw_ds)
```

1 >>

name	gender	height	weight	eyecolor	haircolor	skincolor	homeland	born	died	jedi	species	weapon	friends
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"
"Lando Calrissian"	"male"	1.78	"79"	"brown"	"blank"	"dark"	"Socorro"	"31BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Han Solo"
"Anakin Skywalker"	"male"	1.88	"84"	"blue"	"blond"	"fair"	"Tatooine"	"41.9BBY"	"4ABY"	"jedi"	"human"	"lightsaber"	"Sheev Palpatine"
"Luke Skywalker"	"male"	1.72	"77"	"blue"	"blond"	"fair"	"Tatooine"	"19BBY"	"unk_died"	"jedi"	"human"	"lightsaber"	"Han Solo, Leia Skywalker"
"Leia Skywalker"	"female"	1.5	"49"	"brown"	"brown"	"light"	"Alderaan"	"19BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Obi-Wan Kenobi"
"Obi-Wan Kenobi"	"male"	1.82	"77"	"bluegray"	"auburn"	"fair"	"Stewjon"	"57BBY"	"0BBY"	"jedi"	"human"	"lightsaber"	"Yoda, Qui-Gon Jinn"
"Han Solo"	"male"	1.8	"80"	"brown"	"brown"	"light"	"Corellia"	"29BBY"	"unk_died"	"no_jedi"	"human"	"blaster"	"Leia Skywalker, Luke Skywalker, Obi-Wan Kenobi, Chewbacca"
"Sheev Palpatine"	"male"	1.73	"75"	"blue"	"red"	"pale"	"Naboo"	"82BBY"	"10ABY"	"no_jedi"	"human"	"force-lightning"	"Anakin Skywalker"
"R2-D2"	"male"	0.96	"32"	"NA"	"NA"	"NA"	"Naboo"	"33BBY"	"unk_died"	"no_jedi"	"droid"	"unarmed"	"C-3PO"
"C-3PO"	"male"	1.67	"75"	"NA"	"NA"	"NA"	"Tatooine"	"112BBY"	"3ABY"	"no_jedi"	"droid"	"unarmed"	"R2-D2"
"Yoda"	"male"	0.66	"17"	"brown"	"brown"	"green"	"unk_planet"	"896BBY"	"4ABY"	"jedi"	"yoda"	"lightsaber"	"Obi-Wan Kenobi"
"Darth Maul"	"male"	1.75	"80"	"yellow"	"none"	"red"	"Dathomir"	"54BBY"	"unk_died"	"no_jedi"	"dathomirian"	"lightsaber"	"Sheev Palpatine"
"Chewbacca"	"male"	2.28	"112"	"blue"	"brown"	"NA"	"Kashyyyk"	"200BBY"	"25ABY"	"no_jedi"	"wookiee"	"bowcaster"	"Han Solo"
"Jabba"	"male"	3.9	"NA"	"yellow"	"none"	"tan-green"	"Tatooine"	"unk_born"	"4ABY"	"no_jedi"	"hutt"	"unarmed"	"Boba Fett"

Took: 667 milliseconds, at 2017-6-14 16:30

```
sw_ds.union(sw_ds).show()
```

Took: 712 milliseconds, at 2017-6-14 16:30

Other useful functions

```
sw_ds.count()
```

13

Took: 625 milliseconds, at 2017-6-14 16:31

```
sw_ds.columns.size
```

14

Took: 529 milliseconds, at 2017-6-14 16:31

```
sw_ds.printSchema
```

Took: 521 milliseconds, at 2017-6-14 16:31

```
sw_ds.dtypes
```

14 entries total

Show: 10

Search:

_1	_2
name	StringType
gender	StringType
height	DoubleType
weight	StringType
eyecolor	StringType
haircolor	StringType
skincolor	StringType
homeland	StringType
born	StringType
died	StringType

Showing 1 to 10 of 14 records

Pages: Previous 1 2 Next

Took: 631 milliseconds, at 2017-6-14 16:32

Some more aggregation functions

```
sw_ds.agg(corr($"height", $"weight").as[Double])
```

1

Took: 710 milliseconds, at 2017-6-14 16:33

corr(height, weight)
0.9823964963433255

```
sw_ds.agg(corr($"height", $"weight").as[Double]).show()
```

Took: 662 milliseconds, at 2017-6-14 16:33

```
sw_ds.groupByKey(_.jedi).agg(corr($"height", $"weight").as[Double]).show()
```

Took: 982 milliseconds, at 2017-6-14 16:34

```
sw_ds.groupByKey(_.species).agg(first($"name").as[String]).show()
```

Took: 987 milliseconds, at 2017-6-14 16:34

Other useful functions for creating new columns

```
sw_ds.withColumn("hashed_hair", hash(sw_ds("haircolor"))).show()
```

Took: 596 milliseconds, at 2017-6-14 16:35

```
sw_ds
  .map(x => (x.name, x.friends.split(", ")) )
  .withColumn("NrOfFriendsListed", size($"_2")).show()
```

Took: 744 milliseconds, at 2017-6-14 16:35

```
sw_ds.withColumn("id", monotonically_increasing_id).show()
```

Took: 599 milliseconds, at 2017-6-14 16:36

```
sw_ds.withColumn("random",rand).show()
```

Took: 619 milliseconds, at 2017-6-14 16:36

```
sw_ds.withColumn("name_lenth", length(sw_ds("name"))).show()
```

Took: 608 milliseconds, at 2017-6-14 16:36

```
sw_ds.withColumn("name_species_diff", levenshtein(sw_ds("name"), sw_ds("species"))).show()
```

Took: 583 milliseconds, at 2017-6-14 16:37

```
sw_ds.withColumn("Loc_y", locate("S", sw_ds("name"))).show()
```

Took: 621 milliseconds, at 2017-6-14 16:37

Friendcount example

First Solution

```
sw_ds
  .map(x => x.friends)
  .flatMap(_.split(", "))
  .groupByKey(_.toString)
  .count()
  .show()
```

Took: 1 second 58 milliseconds, at 2017-6-14 16:38

```
ds_missing
  .map(x => x.friends)
  .flatMap(_.getOrElse("").split(", "))
  .groupByKey(_.toString)
  .count()
  .show()
```

Took: 845 milliseconds, at 2017-6-14 16:38

Second Solution

```
import org.apache.spark.sql.functions.explode
```

Took: 391 milliseconds, at 2017-6-14 16:40

```
sw_ds
  .map(x => (x.name, x.friends.split(", ")) )
  .withColumn("friend", explode($"_2"))
  .show()
```

Took: 760 milliseconds, at 2017-6-14 16:40

```
case class NameFriend(name: String, friend: String)
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 405 milliseconds, at 2017-6-14 16:40

```
val NameFriend_df =sw_ds
  .map(x => (x.name, x.friends.split(", ")) )
  .withColumn("_2", explode($"_2"))
  .toDF(Seq("name", "friend"): _*)
```

Took: 494 milliseconds, at 2017-6-14 16:41

```
val NameFriend_ds = NameFriend_df.as[NameFriend]
```

Took: 422 milliseconds, at 2017-6-14 16:41

```
NameFriend_ds.show()
```

Took: 617 milliseconds, at 2017-6-14 16:41

```
NameFriend_ds
  .groupByKey(_.friend)
  .count()
  .orderBy($"count(1)".desc)
  .show()
```

Took: 1 second 15 milliseconds, at 2017-6-14 16:41

```
NameFriend_ds
  .groupByKey(_.friend)
  .count()
  .orderBy($"count(1)".desc)
  .show()
```

Took: 883 milliseconds, at 2017-6-14 16:42

```
case class NameFriendS_ds(name: String, friend: String, S_in_friend:Integer)
org.apache.spark.sql.catalyst.encoders.OuterScopes.addOuterScope(this)
```

Took: 421 milliseconds, at 2017-6-14 16:42

```
NameFriend_ds
 .withColumn("S_in_friend", locate("S", (NameFriend_ds("friend"))) )
 .as[NameFriendS_ds]
 .filter(x=>x.S_in_friend>0)
 .groupByKey(_.name)
 .count()
 .orderBy($"count(1)".desc)
 .show()
```

Took: 1 second 25 milliseconds, at 2017-6-14 16:42

Build: | **buildTime**-*Wed Nov 23 12:19:16 UTC 2016* | **formattedShaVersion**-*0.7.0-c955e71d0204599035f603109527e679aa3bd570* | **sbtVersion**-*0.13.8* | **scalaVersion**-*2.11.8* | **sparkNotebookVersion**-*0.7.0* | **hadoopVersion**-*2.7.2* | **jets3tVersion**-*0.7.1* | **jlineDef**-*(jline,2.12)* | **sparkVersion**-*2.0.2* | **withHive**-*true* |.