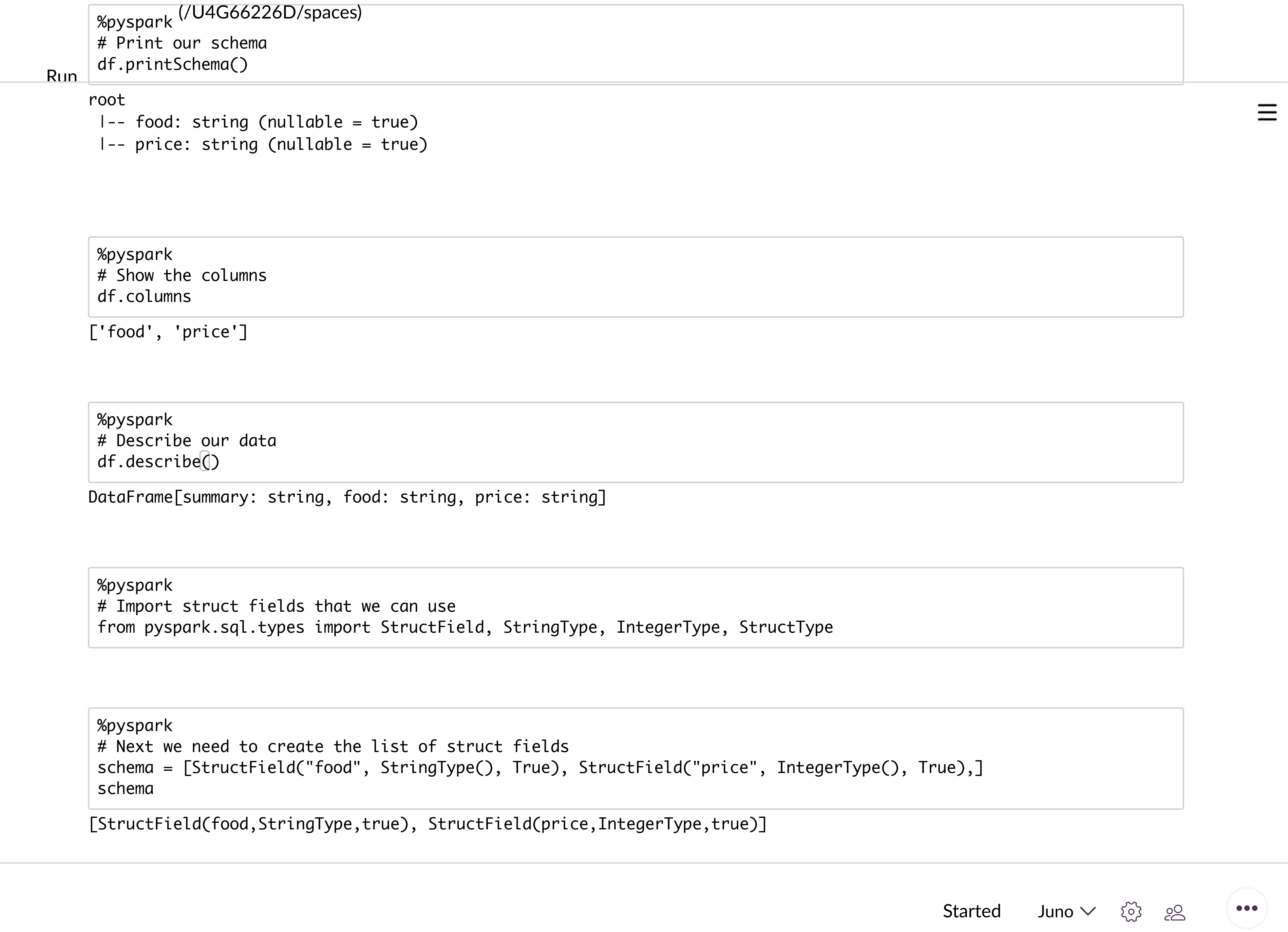
spark_dataframe_basics by ars0107

Dataframe Basics

```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/dataviz-curriculum/day_1/food.csv"
spark.sparkContext.addFile(url)
df = spark.read.csv(SparkFiles.get("food.csv"), sep=",", header=True)

# Show DataFrame
df.show()
```

```
+-------+-----+
|   food|price|
+-------+-----+
|  pizza|    0|
|  sushi|   12|
|chinese|   10|
+-------+-----+
```



```
%pyspark
# Print our schema
df.printSchema()
```

```
root
 |-- food: string (nullable = true)
 |-- price: string (nullable = true)
```

```
%pyspark
# Show the columns
df.columns
```

```
['food', 'price']
```

```
%pyspark
# Describe our data
df.describe()
```

```
DataFrame[summary: string, food: string, price: string]
```

```
%pyspark
# Import struct fields that we can use
from pyspark.sql.types import StructField, StringType, IntegerType, StructType
```

```
%pyspark
# Next we need to create the list of struct fields
schema = [StructField("food", StringType(), True), StructField("price", IntegerType(), True),]
schema
```

```
[StructField(food,StringType,true), StructField(price,IntegerType,true)]
```

```
%pyspark
# Pass in our fields
final = StructType(fields=schema)
final
```

```
StructType(List(StructField(food,StringType,true),StructField(price,IntegerType,true)))
```

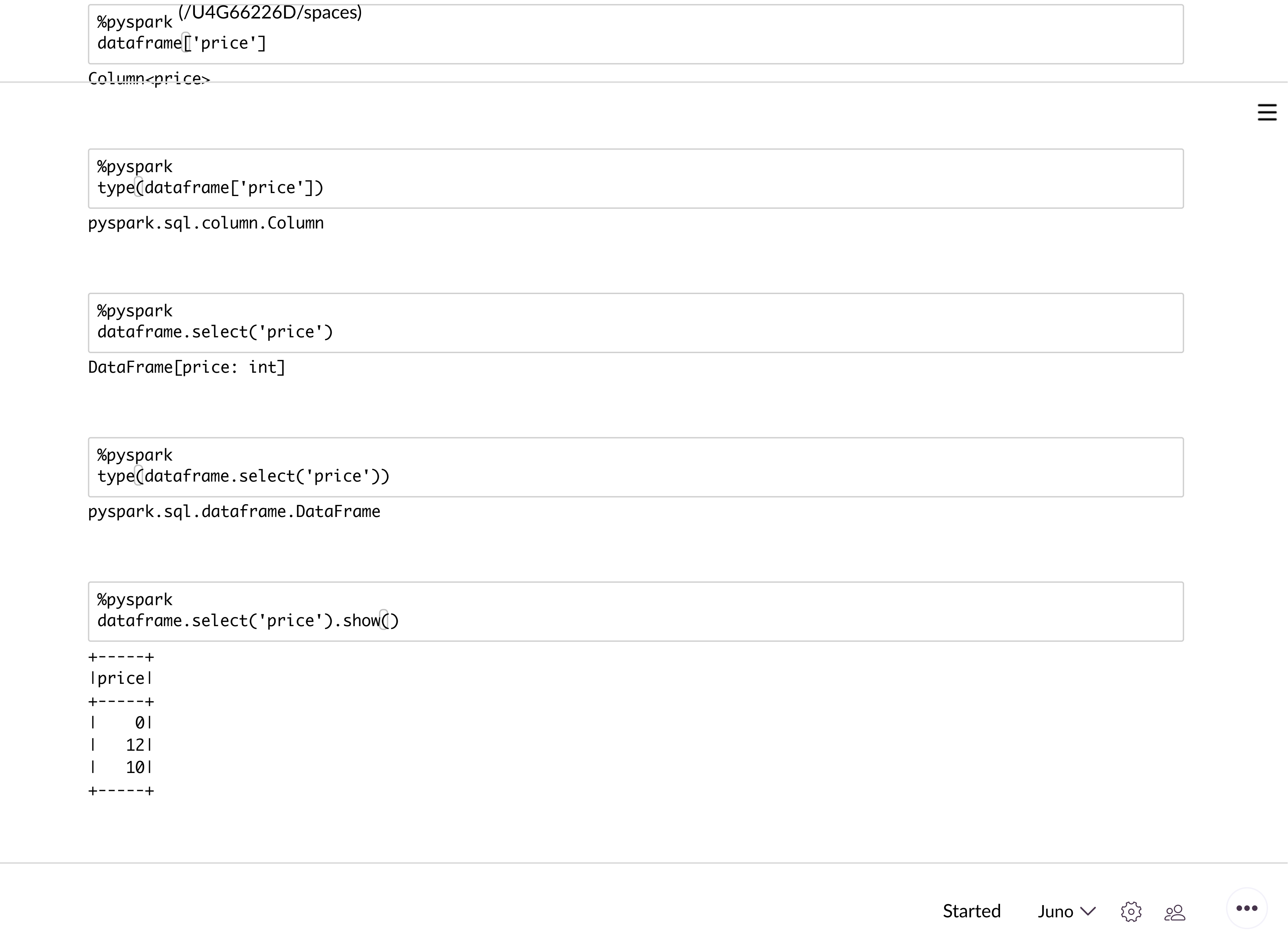
```
%pyspark
# Read our data with our new schema
dataframe = spark.read.csv(SparkFiles.get("food.csv"), schema=final, sep=",", header=True)
dataframe.show()
```

```
+-------+-----+
|   food|price|
+-------+-----+
|  pizza|    0|
|  sushi|   12|
|chinese|   10|
+-------+-----+
```

```
%pyspark
# Print it out
dataframe.printSchema()
```

```
root
 |-- food: string (nullable = true)
 |-- price: integer (nullable = true)
```

Accessing data



(/U4G66226D/spaces)

```
%pyspark
dataframe['price']
```

```
Column<price>
```

```
%pyspark
type(dataframe['price'])
```

```
pyspark.sql.column.Column
```

```
%pyspark
dataframe.select('price')
```

```
DataFrame[price: int]
```

```
%pyspark
type(dataframe.select('price'))
```

```
pyspark.sql.dataframe.DataFrame
```

```
%pyspark
dataframe.select('price').show()
```

```
+-----+
|price|
+-----+
|    0|
|   12|
|   10|
+-----+
```

(/U4G66226D/spaces)

Manipulating Columns

```
%pyspark
# Add new column
dataframe.withColumn('newprice', dataframe['price']).show()
```

```
+-------+-----+--------+
|   food|price|newprice|
+-------+-----+--------+
|  pizza|    0|       0|
|  sushi|   12|      12|
|chinese|   10|      10|
+-------+-----+--------+
```

```
%pyspark
# Update column name
dataframe.withColumnRenamed('price','newerprice').show()
```

```
+-------+----------+
|   food|newerprice|
+-------+----------+
|  pizza|         0|
|  sushi|        12|
|chinese|        10|
+-------+----------+
```

(/U4G66226D/spaces)

```
%pyspark
# Double the price
dataframe.withColumn('doubleprice',dataframe['price']*2).show()
```

```
+-------+-----+-----------+
|   food|price|doubleprice|
+-------+-----+-----------+
|  pizza|    0|          0|
|  sushi|   12|         24|
|chinese|   10|         20|
+-------+-----+-----------+
```

```
%pyspark
# Add a dollar to the price
dataframe.withColumn('add_one_dollar',dataframe['price']+1).show()
```

```
+-------+-----+--------------+
|   food|price|add_one_dollar|
+-------+-----+--------------+
|  pizza|    0|             1|
|  sushi|   12|            13|
|chinese|   10|            11|
+-------+-----+--------------+
```

```
%pyspark
# Half the price
dataframe.withColumn('half_price',dataframe['price']/2).show()
```

```
+-------+-----+----------+
|   food|price|half_price|
+-------+-----+----------+
|  pizza|    0|       0.0|
|  sushi|   12|       6.0|
|chinese|   10|       5.0|
+-------+-----+----------+
```

Started Juno

(/U4G66226D/spaces)

```
%pyspark
# Collecting a column as a list
dataframe.select("price").collect()
```

```
[Row(price=0), Row(price=12), Row(price=10)]
```

Converting PySpark DataFrame to Pandas DataFrame

```
%pyspark
import pandas as pd
pandas_df = dataframe.toPandas()
```

```
%pyspark
pandas_df.head()
```

```
      food  price
0    pizza      0
1    sushi     12
2  chinese     10
```

Interpreter: spark.

Started Juno