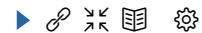


Filtering Data

```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles

url = "https://s3.amazonaws.com/dataviz-curriculum/day_1/wine.csv"
spark.sparkContext.addFile(url)
df = spark.read.csv(SparkFiles.get("wine.csv"), sep=",", header=True)
df.show()
```



id	name	description	year	price	region	subregion	grape
1	Tinta de Toro	Bodega Carmen Rod...	95	110	Oregon	Chehalem Mountains	Willamette V
2	US	This re-named vin...	95	65	Oregon	Chehalem Mountains	Willamette V
3	US	The producer sour...	95	60	California	Sonoma Coast	S
4	Italy	Elegance, complex...	95	80	Northeastern Italy	Colli	
5	US	From 18-year-old ...	95	48	Oregon	Ribbon Ridge	Willamette V
6	US	A standout even i...	95	48	Oregon	Dundee Hills	Willamette V
7	France	This wine is in p...	95	90	Southwest France	Madiran	
8	US	With its sophisti...	95	185	Oregon	Dundee Hills	Willamette V
9	US	First made in 200...	95	90	Oregon	Willamette Valley	Willamette V
10	US	This blockbuster,...	95	325	California	Diamond Mountain	
11	Spain	Nicely oaked blac...	95	80	Northern Spain	Ribera del Duero	
12	France	Coming from a sev...	95	290	Southwest France	Started	Jun
13	France	Coming from a sev...	95	290	Southwest France	Started	Jun

Run



US	This fresh and Li...	Gap's Crown Vineyard	951	751	California	Sonoma Coast	S
onoma	(/U4066240/spaces)	Gary Farrell					

only showing top 20 rows

Interpreter: spark.pyspark. **FINISHED** Took 33 sec 354 millisec. Updated by ars0107 on February 01 2019, 1:09:19 PM (CST)



```
%pyspark
df.printSchema()
```

```
root
 |-- country: string (nullable = true)
 |-- description: string (nullable = true)
 |-- designation: string (nullable = true)
 |-- points: string (nullable = true)
 |-- price: string (nullable = true)
 |-- province: string (nullable = true)
 |-- region_1: string (nullable = true)
 |-- region_2: string (nullable = true)
 |-- variety: string (nullable = true)
 |-- winery: string (nullable = true)
```

Run

Started

Juno ▾



```
%pyspark (/U4G66226D/spaces)
# Order a DataFrame by ascending values
df.orderBy(df["points"].asc()).show(5)
```



```
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+
|country|      description|      designation|points|price|province|      region_1|      region_2|variety|
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+
|    US|This shows a deft...|      null|    null|    null|    null|      null|      null|    null|
|      null|
|    US|The strong aroma ...|      null|    null|    null|    null|      null|      null|    null|
|      null|
| Italy|"This nicely stru...| bitter almond an...|    null|    88|    24|Northeastern Italy|Venezia Giulia|    null|P
inot Grigio|
| Italy|This offers gener...|      null|    null|    null|    null|      null|      null|    null|
|      null|
|    US|      Ripe |      null|    null|    null|    null|      null|      null|    null|
|      null|
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+
only showing top 5 rows
```

```
%pyspark
# Import functions
from pyspark.sql.functions import avg
df.select(avg("points")).show()
```

```
+-----+
|      avg(points)|
+-----+
|87.88834105383143|
+-----+
```

Run

```
%pyspark (/U4G66226D/spaces)
# Using SQL
df.filter("price<20").show()
```



ke Co...	Chardonnay	Parducci	86	10	Douro	null	
Portugal	This family-owned...	Muros de Vinhal	86	10	Douro	null	
	Portuguese Red	Quinta do Portall					
France	This is a smooth,...	Chateau Beauvilla...	86	14	Southwest France	Cahors	
	Malbec-Merlot	Rigall					
US	This is an easygo...	null	86	18	California	California	Californ
Other	Chardonnay	The Naked Grapel					
France	Closer to Bordeaux...	Pigmentum	86	15	Southwest France	Buzet	
	Merlot-Malbec	Georges Vigouroux					
France	This is a blend o...	Pigmentum	86	10	Southwest France	Ctes de Gascogne	
	Ugni Blanc-Colombard	Georges Vigouroux					
US	Aromas of ripe (l...	Tudor Hills Vineyard	86	17	Washington	Yakima Valley	Columbi
a Valley	Pinot Grigio	Martinez & Martinez					
US	Strong wood smoke...	null	86	12	California	Lodi	Centra
l Valley	Cabernet Franc	Ironstonel					
US	This medium-bodie...	null	86	10	California	California	Californ
Other	Chardonnay	Leaping Horsel					
US	This wine is dry ...	null	86	13	California	California	Californ
Other	White Blend	Kitchen Sink					
Portugal	This state-owned ...	Companhia das Lez...	86	12	Tejo	null	
	Portuguese Red	Wines & Winemakers					
Argentina	This copper-tinte...	Terroir @_nico Pi...	86	12	Mendoza Province	Tupungato	
	Roş	Zorzal					
Argentina	Aromas of prune, ...	Reserva	86	15	Mendoza Province	Valle de Uco	
	Malbec	Vialba					

```
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
```

only showing top 20 rows

Run

Started

Juno ▾



```
%pyspark (/U4G66226D/spaces)
# Filter by price on certain columns
df.filter("price<20").select(['points', 'country', 'winery', 'price']).show()
```

```
+-----+-----+-----+-----+
|points|  country|          winery|price|
+-----+-----+-----+-----+
|  90| Bulgaria|    Villa Melnik|  15|
|  90|   Spain|    Don Bernardin|  17|
|  90|    US|      De Loach|  18|
|  91|    US|  Trinity Vineyard|  19|
|  91| Portugal|Adega Cooperativa...|  15|
|  86|    US|    Belle Ambiance|  10|
|  86| Portugal| Adega de Cantanhed|  12|
|  86|    US|      Parducci|  13|
|  86| Portugal|  Quinta do Portal|  10|
|  86| France|        Rigal|  14|
|  86|    US|  The Naked Grape|  18|
|  86| France| Georges Vigouroux|  15|
|  86| France| Georges Vigouroux|  10|
|  86|    US| Martinez & Martinez|  17|
|  86|    US|      Ironstone|  12|
|  86|    US|    Leaping Horse|  10|
|  86|    US|    Kitchen Sink|  13|
|  86| Portugal| Wines & Winemakers|  12|
|  86| Argentina|      Zorzal|  12|
|  86| Argentina|    Vialba|  15|
+-----+-----+-----+-----+
only showing top 20 rows
```

Using Python Comparison Operators

Run

Started

Juno ▾



```
%pyspark (/U4G66226D/spaces)
# Same results only this time using python
df.filter(df["price"] < 200).show()
```



Country	Wine Name	Price	Region	Sub-Region
US	Pinot Noir	95	California	Sonoma Coast
Italy	Pinot Noir	95	Northeastern Italy	Collio
US	Pinot Noir	95	Oregon	Ribbon Ridge
US	Pinot Noir	95	Oregon	Dundee Hills
France	Pinot Noir	95	Southwest France	Madiran
US	Pinot Noir	95	Oregon	Dundee Hills
US	Pinot Noir	95	Oregon	Willamette Valley
Spain	Pinot Noir	95	Northern Spain	Ribera del Duero
US	Pinot Noir	95	California	Sonoma Coast
US	Pinot Noir	95	California	Napa Valley
Spain	Pinot Noir	95	Northern Spain	Rioja
US	Pinot Noir	95	California	Edna Valley

only showing top 20 rows

Run

Started

Juno ▾



三


```
%pyspark (/U4G66226D/spaces)
df.filter(df["country"] == "US").show()
```



country	description	designation	points	price	province	region_1	region_2
variety	winery						
US	This tremendous 1...	Martha's Vineyard	96	235	California	Napa Valley	Napa
US	Mac Watson honors...	Special Selected ...	96	90	California	Knights Valley	Sonoma
US	This spent 20 mon...	Reserve	96	65	Oregon	Willamette Valley	Willamette Valley
US	This re-named vin...	Silice	95	65	Oregon	Chehalem Mountains	Willamette Valley
US	The producer sour...	Gap's Crown Vineyard	95	60	California	Sonoma Coast	Sonoma
US	From 18-year-old ...	Estate Vineyard W...	95	48	Oregon	Ribbon Ridge	Willamette Valley
US	A standout even i...	Weber Vineyard	95	48	Oregon	Dundee Hills	Willamette Valley
US	With its sophisti...	Grace Vineyard	95	185	Oregon	Dundee Hills	Willamette Valley
US	First made in 200...	Sigrid	95	90	Oregon	Willamette Valley	Willamette Valley
US	This blockbuster,...	Rainin Vineyard	95	325	California	Diamond Mountain ...	Napa
US	This fresh and li...	Gap's Crown Vineyard	95	75	California	Sonoma Coast	Sonoma
US	Heitz has made th...	Grignolino	95	24	California	Napa Valley	Napa

Run

Started

Juno ▾

