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()
```

```
| Spain|Lush cedary black...|Carodorum Selecci...|    95|  110|    Northern Spain|                Toro|
null|     Tinta de Toro|Bodega Carmen Rod...|
|    US|This re-named vin...|             Silice|    95|   65|            Oregon|  Chehalem Mountains|Willamette V
alley|        Pinot Noir|           Bergström|
|    US|The producer sour...|Gap's Crown Vineyard|    95|   60|        California|        Sonoma Coast|           S
onoma|        Pinot Noir|           Blue Farm|
|  Italy|Elegance, complex...|  Ronco della Chiesa|    95|   80|Northeastern Italy|              Collio|
null|          Friulano|    Borgo del Tiglio|
|    US|From 18-year-old ...|Estate Vineyard W...|    95|   48|            Oregon|        Ribbon Ridge|Willamette V
alley|        Pinot Noir|Patricia Green Ce...|
|    US|A standout even i...|      Weber Vineyard|    95|   48|            Oregon|        Dundee Hills|Willamette V
alley|        Pinot Noir|Patricia Green Ce...|
| France|This wine is in p...|Château Montus Pr...|    95|   90|  Southwest France|             Madiran|
null|           Tannat|   Vignobles Brumont|
|    US|With its sophisti...|      Grace Vineyard|    95|  185|            Oregon|        Dundee Hills|Willamette V
alley|        Pinot Noir|      Domaine Serene|
|    US|First made in 200...|             Sigrid|    95|   90|            Oregon|   Willamette Valley|Willamette V
alley|        Chardonnay|           Bergström|
|    US|This blockbuster,...|     Rainin Vineyard|    95|  325|        California|Diamond Mountain ...|
Napa|Cabernet Sauvignon|               Hall|
|  Spain|Nicely oaked blac...|6 Años Reserva Pr...|    95|   80|    Northern Spain|    Ribera del Duero|
null|       Tempranillo|           Valduero|
| France|Coming from a sev...|       Le Pigeonnier|    95|  290|  Southwest France|              Cahors|
null|           Malbec|  Château Lagrézette|
```

```
|     US|This fresh and li...|Gap's Crown Vineyard|    95|   75|        California|       Sonoma Coast|           S
onoma|    (/U4C66226D/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|
     winery|
+-------+--------------------+--------------------+------+-----+--------+-----------------+-------------+-------+-
-----------+
|     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

Started Juno

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

```
+--------+--------------------+--------------------+------+-----+----------------+--------------------+------------
--------+--------------------+--------------------+
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 Bernardino|   17|
|    90|       US|             De Loach|   18|
|    91|       US|   Trinity Vineyards|   19|
|    91| Portugal|Adega Cooperativa...|   15|
|    86|       US|      Belle Ambiance|   10|
|    86| Portugal| Adega de Cantanhede|   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|             Vi̱alba|   15|
+------+---------+--------------------+-----+
only showing top 20 rows
```

Using Python Comparison Operators

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

```
+-------+--------------------+--------------------+------+-----+-----------------+-----------------+--------------
---+-----------------+--------------------+
only showing top 20 rows
```

```
%pyspark (/U4G66226D/spaces)
df.filter( (df["price"] < 200) | (df['points'] > 80) ).show()
```

```
| Spain|Lush cedary black...|Carodorum Selecci...|    95|  110|    Northern Spain|                Toro|
null|     Tinta de Toro|Bodega Carmen Rod...|
|    US|This re-named vin...|             Silice|    95|   65|            Oregon|  Chehalem Mountains|Willamette V
alley|        Pinot Noir|          Bergström|
|    US|The producer sour...|Gap's Crown Vineyard|    95|   60|        California|        Sonoma Coast|           S
onoma|        Pinot Noir|          Blue Farm|
|  Italy|Elegance, complex...|  Ronco della Chiesa|    95|   80|Northeastern Italy|              Collio|
null|          Friulano|    Borgo del Tiglio|
|    US|From 18-year-old ...|Estate Vineyard W...|    95|   48|            Oregon|        Ribbon Ridge|Willamette V
alley|        Pinot Noir|Patricia Green Ce...|
|    US|A standout even i...|      Weber Vineyard|    95|   48|            Oregon|        Dundee Hills|Willamette V
alley|        Pinot Noir|Patricia Green Ce...|
| France|This wine is in p...|Château Montus Pr...|    95|   90|  Southwest France|             Madiran|
null|            Tannat|   Vignobles Brumont|
|    US|With its sophisti...|      Grace Vineyard|    95|  185|            Oregon|        Dundee Hills|Willamette V
alley|        Pinot Noir|      Domaine Serene|
|    US|First made in 200...|             Sigrid|    95|   90|            Oregon|   Willamette Valley|Willamette V
alley|        Chardonnay|          Bergström|
|    US|This blockbuster,...|     Rainin Vineyard|    95|  325|        California|Diamond Mountain ...|
Napa|Cabernet Sauvignon|               Hall|
|  Spain|Nicely oaked blac...|6 Años Reserva Pr...|    95|   80|    Northern Spain|    Ribera del Duero|
null|       Tempranillo|           Valduero|
| France|Coming from a sev...|       Le Pigeonnier|    95|  290|  Southwest France|              Cahors|
null|            Malbec|  Château Lagrzette|
|    US|This fresh and li...|Gap's Crown Vineyard|    95|   75|        California|        Sonoma Coast|           S
onoma|        Pinot Noir|        Gary Farrell|
+-------+--------------------+--------------------+------+-----+------------------+--------------------+------------
-----+------------------+--------------------+
only showing top 20 rows
```

```
%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|Ca
bernet Sauvignon|              Heitz|
|     US|Mac Watson honors...|Special Selected ...|    96|   90|California|      Knights Valley|          Sonoma|
Sauvignon Blanc|           Macauley|
|     US|This spent 20 mon...|             Reserve|    96|   65|    Oregon|   Willamette Valley|Willamette Valley|
      Pinot Noir|              Ponzi|
|     US|This re-named vin...|              Silice|    95|   65|    Oregon|  Chehalem Mountains|Willamette Valley|
      Pinot Noir|          Bergström|
|     US|The producer sour...|Gap's Crown Vineyard|    95|   60|California|        Sonoma Coast|          Sonoma|
      Pinot Noir|           Blue Farm|
|     US|From 18-year-old ...|Estate Vineyard W...|    95|   48|    Oregon|        Ribbon Ridge|Willamette Valley|
      Pinot Noir|Patricia Green Ce...|
|     US|A standout even i...|      Weber Vineyard|    95|   48|    Oregon|        Dundee Hills|Willamette Valley|
      Pinot Noir|Patricia Green Ce...|
|     US|With its sophisti...|      Grace Vineyard|    95|  185|    Oregon|        Dundee Hills|Willamette Valley|
      Pinot Noir|      Domaine Serene|
|     US|First made in 200...|              Sigrid|    95|   90|    Oregon|   Willamette Valley|Willamette Valley|
      Chardonnay|          Bergström|
|     US|This blockbuster,...|     Rainin Vineyard|    95|  325|California|Diamond Mountain ...|            Napa|Ca
bernet Sauvignon|               Hall|
|     US|This fresh and li...|Gap's Crown Vineyard|    95|   75|California|        Sonoma Coast|          Sonoma|
      Pinot Noir|        Gary Farrell|
|     US|Heitz has made th...|          Grignolino|    95|   24|California|         Napa Valley|            Napa|
```

Run

Started

Juno