PySpark Crash Course

January 27, 2021

```
[5]: # Load Pyspark Pkgs
     import pyspark
[7]: # Methods/Attrib
     dir(pyspark)
[7]: ['Accumulator',
      'AccumulatorParam',
      'BarrierTaskContext',
      'BarrierTaskInfo',
      'BasicProfiler',
      'Broadcast',
      'HiveContext',
      'MarshalSerializer',
      'PickleSerializer',
      'Profiler',
      'RDD',
      'RDDBarrier',
      'Row',
      'SQLContext',
      'SparkConf',
      'SparkContext',
      'SparkFiles',
      'SparkJobInfo',
      'SparkStageInfo',
      'StatusTracker',
      'StorageLevel',
      'TaskContext',
      '_NoValue',
      '__all__',
      '__builtins__',
      '\_\_cached\_\_',
      '__doc__',
      '__file__',
      '__loader__',
      '__name__',
      '__package__',
```

```
'__path__',
'__spec__',
'__version__',
'_globals',
'accumulators',
'broadcast',
'cloudpickle',
'conf',
'context',
'copy_func',
'files',
'find_spark_home',
'heapq3',
'java_gateway',
'join',
'keyword_only',
'profiler',
'rdd',
'rddsampler',
'resource',
'resultiterable',
'serializers',
'shuffle',
'since',
'sql',
'statcounter',
'status',
'storagelevel',
'taskcontext',
'traceback_utils',
'types',
'util',
'version',
'wraps']
```

Working with DataFrames in PySpark

- Read DataSet(CSV)
- Create DataFrame

Tips

- SparkSession
- SparkContext :sc
- SqlContext

```
[8]: # Create A SparkSession
from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder.appName("PySparkTut").getOrCreate()
 [9]: !ls
     'PySpark Crash Course.ipynb'
[10]: | wget https://raw.githubusercontent.com/Jcharis/common_ml_datasets_explorer_app/
       →master/datasets/diamonds.csv
     --2021-01-27 21:01:43-- https://raw.githubusercontent.com/Jcharis/common_ml_dat
     asets explorer app/master/datasets/diamonds.csv
     Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
     151.101.112.133
     Connecting to raw.githubusercontent.com
     (raw.githubusercontent.com) | 151.101.112.133 | :443... connected.
     HTTP request sent, awaiting response... 200 OK
     Length: 2772143 (2.6M) [text/plain]
     Saving to: 'diamonds.csv'
     diamonds.csv
                          100%[========>]
                                                        2.64M 3.81MB/s
                                                                            in 0.7s
     2021-01-27 21:01:44 (3.81 MB/s) - 'diamonds.csv' saved [2772143/2772143]
[11]:
     !ls
      diamonds.csv 'PySpark Crash Course.ipynb'
[12]: # Read A DataSet without header
      df = spark.read.csv('diamonds.csv')
[13]: # Preview dataset
      df.show()
                                 _c3|
                                       _c4|
        _c0|
                                             _c5| _c6| _c7| _c8| _c9|
                   _c1|
                         _c2|
                  cut|color|clarity|depth|table|price|
     |carat|
                                                           x
     0.23
                Ideal|
                                 SI2| 61.5|
                                              55| 326|3.95|3.98|2.43|
                           E \mid
     | 0.21| Premium|
                           Εl
                                 SI1| 59.8|
                                              61|
                                                   326|3.89|3.84|2.31|
     1 0.231
                  Goodl
                           Εl
                                 VS1| 56.9|
                                              65 l
                                                   327 | 4.05 | 4.07 | 2.31 |
     0.29 Premium
                           Ιl
                                 VS2 | 62.4
                                              58 l
                                                   334 | 4.2 | 4.23 | 2.63 |
     0.31
                 Good
                           J|
                                 SI2| 63.3|
                                              58|
                                                   335 | 4.34 | 4.35 | 2.75 |
     | 0.24|Very Good|
                           J|
                               VVS2| 62.8|
                                              57|
                                                   336|3.94|3.96|2.48|
     | 0.24|Very Good|
                           Ιl
                               VVS1| 62.3|
                                                   336|3.95|3.98|2.47|
                                              57|
     0.26 | Very Good
                          Η|
                                 SI1 | 61.9 |
                                              55 | 337 | 4.07 | 4.11 | 2.53 |
     0.22
                 Fair|
                                 VS2| 65.1|
                           Εl
                                              61|
                                                   337|3.87|3.78|2.49|
     | 0.23|Very Good|
                           НΙ
                                 VS1 | 59.4
                                              61 | 338 | 4 | 4 . 05 | 2 . 39 |
```

```
0.31
            Good
                     JΙ
                           SI1|
                                  64|
                                         55 l
                                              339|4.25|4.28|2.73|
0.23
           Ideal
                     JΙ
                           VS1 | 62.8|
                                         56|
                                              340|3.93| 3.9|2.46|
0.221
         Premium |
                     FΙ
                           SI1 | 60.4|
                                              342|3.88|3.84|2.33|
                                         61|
0.31
                           SI2| 62.2|
                                         54|
                                              344 | 4.35 | 4.37 | 2.71 |
           Ideal|
                     J|
0.21
                                         621
         Premium|
                     Εl
                           SI2 | 60.2 |
                                              345|3.79|3.75|2.27|
0.32
         Premium |
                            I1| 60.9|
                                         58 l
                                              345 | 4.38 | 4.42 | 2.68 |
0.3
           Ideal|
                     Ιl
                           SI2
                                   62|
                                         54|
                                              348 | 4.31 | 4.34 | 2.68 |
0.31
            Good
                     JΙ
                           SI1| 63.4|
                                         54 l
                                              351 | 4.23 | 4.29 | 2.7 |
                                              351 | 4.23 | 4.26 | 2.71 |
0.3
            Good
                     J|
                           SI1| 63.8|
                                         56|
+----+----+----+----+----+
```

only showing top 20 rows

```
[62]: # Read A DataSet with header/column names

df = spark.read.csv('diamonds.csv',header=True)
```

[15]: df.show()

```
---+----+----+----+
              cut|color|clarity|depth|table|price|
                                                             yΙ
0.23
           Ideal|
                      Εl
                             SI2| 61.5|
                                           55 l
                                                326|3.95|3.98|2.43|
0.21
         Premium |
                             SI1| 59.8|
                                                326|3.89|3.84|2.31|
                      Εl
                                           61|
0.23
            Good
                      E \mid
                             VS1 | 56.9|
                                           65 l
                                                327 | 4.05 | 4.07 | 2.31 |
0.29
         Premium |
                      Ιl
                             VS2| 62.4|
                                           58|
                                                334| 4.2|4.23|2.63|
0.31
             Good
                                                335|4.34|4.35|2.75|
                      JΙ
                             SI2| 63.3|
                                           58|
| 0.24|Very Good|
                      J|
                            VVS2| 62.8|
                                           57|
                                                336|3.94|3.96|2.48|
| 0.24|Very Good|
                      Ιl
                            VVS1 | 62.3
                                           57|
                                                336|3.95|3.98|2.47|
| 0.26|Very Good|
                      Η|
                             SI1 | 61.9 |
                                           55|
                                                337 | 4.07 | 4.11 | 2.53 |
0.22
                      Εl
                             VS2| 65.1|
                                                337|3.87|3.78|2.49|
             Fair
                                           61|
| 0.23|Very Good|
                      Η|
                             VS1| 59.4|
                                           61|
                                                338|
                                                        4|4.05|2.39|
0.3
                                                339 | 4.25 | 4.28 | 2.73 |
            Good
                      J|
                             SI1
                                    64 l
                                           55 l
0.23
                             VS1 | 62.8|
                                           56|
                                                340|3.93| 3.9|2.46|
           Ideal|
                      J|
0.22
         Premium|
                             SI1 | 60.4 |
                                           61 l
                                                342|3.88|3.84|2.33|
0.31
                             SI2| 62.2|
                                           54 l
                                                344 | 4.35 | 4.37 | 2.71 |
           Ideal|
                      J|
0.21
         Premium
                      Εl
                             SI2 | 60.2 |
                                           62 l
                                                345 | 3.79 | 3.75 | 2.27 |
1 0.321
         Premium|
                      Εl
                              I1 | 60.9
                                           58 l
                                                345 | 4.38 | 4.42 | 2.68 |
0.3
           Ideal
                                                348 | 4.31 | 4.34 | 2.68 |
                      Ιl
                             SI2|
                                    62|
                                           54|
0.31
            Good
                      J|
                             SI1| 63.4|
                                           54 l
                                                351|4.23|4.29| 2.7|
0.3
                                                351 | 4.23 | 4.26 | 2.71 |
             Good
                      J|
                             SI1| 63.8|
                                           56|
  0.3|Very Good|
                                                351 | 4.21 | 4.27 | 2.66 |
                      JΙ
                             SI1 | 62.7 |
                                           59 l
+----+-
```

```
[16]: # Columns df.columns
```

```
[16]: ['carat', 'cut', 'color', 'clarity', 'depth', 'table', 'price', 'x', 'y', 'z']
[20]: # Shape (rows + columns)
     (df.count() ,len(df.columns))
[20]: (53940, 10)
[21]: # Number of columns
     len(df.columns)
[21]: 10
[22]: # Number of rows
     df.count()
[22]: 53940
[24]: # Descriptive Analysis
     df.describe().show()
     ----+
     |summary|
                                     cut|color|clarity|
                                                                   depth|
                          carat
     table|
                      price
                                            χl
                                                              уl
     z١
     | count|
                          53940|
                                   53940|53940| 53940|
                                                                   539401
     539401
                      539401
                                        53940 l
                                                          539401
     539401
        mean | 0.7979397478679852 |
                                    null| null|
                                                 null| 61.74940489432624|
     57.45718390804603|3932.799721913237| 5.731157211716609|
     5.734525954764462|3.5387337782723316|
                                    null | null | null | 1.4326213188336525 | 2.2344905
     | stddev|0.4740112444054196|
     628213247 | 3989 . 439738146397 | 1 . 1217607467924915 | 1 . 1421346741235616 | 0 . 705698846949
     9883 l
     1
         minl
                            0.21
                                    Fairl
                                             DΙ
                                                   I1
                                                                      431
     431
                                                           01
                    1000
                                         01
                                                                             01
     1
                           5.01|Very Good|
                                             J|
                                                 VVS2|
                                                                      791
         max
     95 l
                    99991
                                                        9.941
                                                                          8.061
                                      9.861
```

```
[25]: # Pick a column & Get summary/describe a selected column
      df.describe('carat').show()
     |summary|
                           carat
       count
                           53940
         mean | 0.7979397478679852 |
     | stddev|0.4740112444054196|
          min
          max |
                            5.01
[26]: # Preview the First Row
      df.first()
[26]: Row(carat='0.23', cut='Ideal', color='E', clarity='SI2', depth='61.5',
      table='55', price='326', x='3.95', y='3.98', z='2.43')
[31]: # Preview the first 10 rows
      # Like a list
      df.head(10)
[31]: [Row(carat='0.23', cut='Ideal', color='E', clarity='SI2', depth='61.5',
      table='55', price='326', x='3.95', y='3.98', z='2.43'),
      Row(carat='0.21', cut='Premium', color='E', clarity='SI1', depth='59.8',
      table='61', price='326', x='3.89', y='3.84', z='2.31'),
       Row(carat='0.23', cut='Good', color='E', clarity='VS1', depth='56.9',
      table='65', price='327', x='4.05', y='4.07', z='2.31'),
      Row(carat='0.29', cut='Premium', color='I', clarity='VS2', depth='62.4',
      table='58', price='334', x='4.2', y='4.23', z='2.63'),
       Row(carat='0.31', cut='Good', color='J', clarity='SI2', depth='63.3',
      table='58', price='335', x='4.34', y='4.35', z='2.75'),
       Row(carat='0.24', cut='Very Good', color='J', clarity='VVS2', depth='62.8',
      table='57', price='336', x='3.94', y='3.96', z='2.48'),
       Row(carat='0.24', cut='Very Good', color='I', clarity='VVS1', depth='62.3',
      table='57', price='336', x='3.95', y='3.98', z='2.47'),
      Row(carat='0.26', cut='Very Good', color='H', clarity='SI1', depth='61.9',
      table='55', price='337', x='4.07', y='4.11', z='2.53'),
      Row(carat='0.22', cut='Fair', color='E', clarity='VS2', depth='65.1',
      table='61', price='337', x='3.87', y='3.78', z='2.49'),
      Row(carat='0.23', cut='Very Good', color='H', clarity='VS1', depth='59.4',
      table='61', price='338', x='4', y='4.05', z='2.39')]
[32]: # Method 2: Useful Action with show()
      # Show first 10 datapoints
```

```
cut|color|clarity|depth|table|price|
     0.23
                          Εl
                                 SI2| 61.5|
                                              55 l
                                                   326|3.95|3.98|2.43|
                Ideal
     0.21
              Premium |
                          Εl
                                SI1| 59.8|
                                              61|
                                                   326|3.89|3.84|2.31|
     0.231
                 Good
                          Εl
                                VS1| 56.9|
                                              65 l
                                                  327 | 4.05 | 4.07 | 2.31 |
     0.291
              Premium
                          Ιl
                                VS2| 62.4|
                                              58|
                                                  334 | 4.2 | 4.23 | 2.63 |
     0.31
                 Good
                          J|
                                SI2| 63.3|
                                              58|
                                                  335|4.34|4.35|2.75|
     | 0.24|Very Good|
                                             57|
                                                  336|3.94|3.96|2.48|
                          J|
                               VVS2| 62.8|
     | 0.24|Very Good|
                              VVS1| 62.3|
                                                  336|3.95|3.98|2.47|
                          Ι|
                                              57|
     | 0.26|Very Good|
                          НΙ
                                SI1| 61.9|
                                              55 l
                                                  337 | 4.07 | 4.11 | 2.53 |
     0.22
                 Fair|
                          Εl
                                VS2| 65.1|
                                              61 | 337 | 3.87 | 3.78 | 2.49 |
     0.23 | Very Good
                          НΙ
                                VS1 | 59.4
                                              61 l
                                                  338 l
                                                         4|4.05|2.39|
     +----+----+----+----+----+
     only showing top 10 rows
[33]: # Get Last Rows
      df.tail(5)
[33]: [Row(carat='0.72', cut='Ideal', color='D', clarity='SI1', depth='60.8',
      table='57', price='2757', x='5.75', y='5.76', z='3.5'),
       Row(carat='0.72', cut='Good', color='D', clarity='SI1', depth='63.1',
      table='55', price='2757', x='5.69', y='5.75', z='3.61'),
       Row(carat='0.7', cut='Very Good', color='D', clarity='SI1', depth='62.8',
      table='60', price='2757', x='5.66', y='5.68', z='3.56'),
       Row(carat='0.86', cut='Premium', color='H', clarity='SI2', depth='61',
      table='58', price='2757', x='6.15', y='6.12', z='3.74'),
      Row(carat='0.75', cut='Ideal', color='D', clarity='SI2', depth='62.2',
      table='55', price='2757', x='5.83', y='5.87', z='3.64')]
     0.0.1 Selection of columns
        • .select ###### Note
        • Dot & Bracket Notation only gives the column name not the entire column
            - ['colA']*
            - .colA*
[35]: # List all Columns
      df.columns
[35]: ['carat', 'cut', 'color', 'clarity', 'depth', 'table', 'price', 'x', 'y', 'z']
[37]: # Select A Column
      df.select('carat').show()
```

df.show(10)

```
+---+
    |carat|
    +---+
    0.23
    0.21
    0.23
    0.29
    0.31
    0.24
    0.24
    0.26
    0.22
    0.23
    0.31
    0.23
    0.221
    | 0.31|
    0.2
    0.32
    0.3
    0.3
    0.3
    0.3
    +---+
    only showing top 20 rows
[40]: # Select A Column irrespective of column word case
     # will work irrespective of the case of the column once it is found within the
     \rightarrow dataset
     df.select('CARAT').show()
    +----+
    |CARAT|
    +----+
    0.23
    0.21
    0.23
    0.29
    0.31
    0.24
    0.24
    0.26
    0.22
    0.23
    0.3
    0.23
    0.22
```

```
0.31
    0.21
    0.321
    0.3
    1 0.31
    0.3
    0.3
    +---+
    only showing top 20 rows
[41]: # This is not as we would expect in pandas
     # For Bracket Notation : pick column name not the entire columne
     df['carat']
[41]: Column<br/>carat'>
[44]: # This is not as we would expect in pandas
     # For Dot Notation : pick column name not the entire column
     df.carat
[44]: Column<br/>carat'>
[45]: # Select Multiple Columns
     df.select('carat','cut').show(5)
    +----+
    |carat|
             cutl
    | 0.23| Ideal|
    | 0.21|Premium|
    0.23
             Good
    | 0.29|Premium|
    | 0.31| Good|
    +----+
    only showing top 5 rows
    0.0.2 Column Filtering and Applying Conditions
       • .filter
       • .where
[46]: # Filter of Columns
     # Apply A Condition
     df.show(10)
    cut|color|clarity|depth|table|price| x| y|
     |carat|
```

```
0.23
           Ideal
                     Εl
                           SI2| 61.5|
                                        55|
                                             326 | 3.95 | 3.98 | 2.43 |
0.21
        Premium |
                     Εl
                           SI1| 59.8|
                                             326|3.89|3.84|2.31|
                                        61|
0.23
            Good
                          VS1| 56.9|
                                             327 | 4.05 | 4.07 | 2.31 |
                     E \mid
                                        65|
0.29
        Premium|
                     Ιl
                          VS2 | 62.4 |
                                        58 l
                                             334 | 4.2 | 4.23 | 2.63 |
0.31
           Good
                                             335|4.34|4.35|2.75|
                     J|
                           SI2| 63.3|
                                        58 l
| 0.24|Very Good|
                     J|
                         VVS2| 62.8|
                                        57|
                                             336|3.94|3.96|2.48|
| 0.24|Very Good|
                     Ιl
                         VVS1 | 62.3 |
                                        57 l
                                             336|3.95|3.98|2.47|
| 0.26|Very Good|
                     Η|
                           SI1 | 61.9 |
                                        55|
                                            337 | 4.07 | 4.11 | 2.53 |
0.22
           Fair
                     E \mid
                          VS2| 65.1|
                                        61|
                                             337|3.87|3.78|2.49|
| 0.23|Very Good|
                           VS1| 59.4|
                                             338|
                                                   4|4.05|2.39|
                     Η|
                                        61|
```

only showing top 10 rows

```
[47]: # Method 1:using filter
df.filter(df['cut'] == "Good").show()
```

```
+----+
|carat| cut|color|clarity|depth|table|price|
                                                         yΙ
+----+---+----+----+
| 0.23|Good|
                 Εl
                        VS1| 56.9|
                                      65|
                                            327 | 4.05 | 4.07 | 2.31 |
| 0.31|Good|
                 J|
                                            335 | 4.34 | 4.35 | 2.75 |
                        SI2| 63.3|
                                      58 l
| 0.3|Good|
                 JΙ
                        SI1|
                                64|
                                      55 l
                                            339 | 4.25 | 4.28 | 2.73 |
| 0.3|Good|
                        SI1| 63.4|
                                            351 | 4.23 | 4.29 | 2.7 |
                 J|
                                      54|
| 0.3|Good|
                        SI1| 63.8|
                                      56|
                                            351 | 4.23 | 4.26 | 2.71 |
                 J
| 0.3|Good|
                        SI2| 63.3|
                 Ιl
                                      56|
                                            351|4.26| 4.3|2.71|
| 0.23|Good|
                        VS1 | 58.2
                                            402|4.06|4.08|2.37|
                 FΙ
                                      59|
| 0.23|Good|
                 E|
                        VS1 | 64.1|
                                      59|
                                            402|3.83|3.85|2.46|
                 Ηl
| 0.31|Good|
                        SI11
                                      54 l
                                            402|4.29|4.31|2.75|
                                64 l
| 0.26|Good|
                 DΙ
                        VS2| 65.2|
                                      561
                                            403|3.99|4.02|2.61|
| 0.26|Good|
                 DΙ
                        VS1 | 58.4 |
                                            403 | 4.19 | 4.24 | 2.46 |
                                      63 l
| 0.32|Good|
                        SI2| 63.1|
                                            403 | 4.34 | 4.37 | 2.75 |
                 Η|
                                      561
| 0.32|Good|
                 H \mid
                        SI2 | 63.8 |
                                      56 l
                                            403 | 4.36 | 4.38 | 2.79 |
| 0.3|Good|
                 Ιl
                        SI1| 63.2|
                                      55|
                                            405 | 4.25 | 4.29 | 2.7 |
| 0.3|Good|
                 H \mid
                        SI1 | 63.7 |
                                      57 l
                                            554 | 4.28 | 4.26 | 2.72 |
| 0.26|Good|
                 Εl
                       VVS1| 57.9|
                                      60 l
                                            554|4.22|4.25|2.45|
| 0.7|Good|
                                      58 | 2759 | 5.85 | 5.9 | 3.38 |
                 E \mid
                        VS2 | 57.5 |
| 0.7|Good|
                 FΙ
                        VS1 | 59.4|
                                      62 | 2759 | 5.71 | 5.76 | 3.4 |
| 0.7|Good|
                       VVS2| 62.1|
                                      64 | 2767 | 5.62 | 5.65 | 3.5 |
                 Η|
| 0.71|Good|
                 E|
                        VS2 | 59.2
                                      61 | 2772 | 5.8 | 5.88 | 3.46 |
```

```
[48]: # Method 1:using filter
df.filter(df.carat >= 0.7).show()
```

```
cut|color|clarity|depth|table|price|
|carat|
                                                           x
                                                                 уl
0.7
                        Εl
                               SI1| 62.5|
                                             57 | 2757 | 5.7 | 5.7 | 3.57 |
            Ideal
0.86
             Fair
                       Εl
                              SI2 | 55.1 |
                                             69 | 2757 | 6.45 | 6.33 | 3.52 |
   0.7|
            Ideal|
                        G|
                              VS2 | 61.6|
                                             56 | 2757 | 5.7 | 5.67 | 3.5 |
| 0.71|Very Good|
                       E|
                              VS2 | 62.4 |
                                             57 | 2759 | 5.68 | 5.73 | 3.56 |
| 0.78|Very Good|
                       G|
                              SI2 | 63.8 |
                                             56 | 2759 | 5.81 | 5.85 | 3.72 |
  0.7
             Good
                       Εl
                              VS2 | 57.5 |
                                             58 | 2759 | 5.85 | 5.9 | 3.38 |
   0.71
             Good
                       FΙ
                              VS1 | 59.4 |
                                             62 | 2759 | 5.71 | 5.76 | 3.4 |
0.96
                       FΙ
                              SI2| 66.3|
                                             62 | 2759 | 6.27 | 5.95 | 4.07 |
             Fair|
| 0.73|Very Good|
                        Εl
                              SI1 | 61.6|
                                             59 | 2760 | 5.77 | 5.78 | 3.56 |
          Premium |
   0.8
                              SI1 | 61.5 |
                                             58 | 2760 | 5.97 | 5.93 | 3.66 |
| 0.75|Very Good|
                       DΙ
                              SI1 63.2
                                             56 | 2760 | 5.8 | 5.75 | 3.65 |
0.75
          Premium |
                       Εl
                              SI1| 59.9|
                                             54 | 2760 |
                                                           6|5.96|3.58|
0.74
            Ideal
                        G|
                              SI1 | 61.6|
                                             55 | 2760 | 5.8 | 5.85 | 3.59 |
0.75
          Premium |
                        G|
                              VS2 | 61.7 |
                                             58 | 2760 | 5.85 | 5.79 | 3.59 |
1 0.81
                              VS1 | 62.9 |
                                             56 | 2760 | 5.94 | 5.87 | 3.72 |
            Ideal|
                        Ιl
| 0.75|
            Ideal|
                        G|
                              SI1 | 62.2 |
                                             55 | 2760 | 5.87 | 5.8 | 3.63 |
  0.81
                        GΙ
                                             59 | 2760 | 5.9 | 5.81 | 3.69 |
          Premium |
                              SI1
                                      63 l
| 0.74|
            Ideal|
                        Ιl
                             VVS2| 62.3|
                                             55 | 2761 | 5.77 | 5.81 | 3.61 |
0.81
            Ideal|
                       FΙ
                              SI2| 58.8|
                                             57 | 2761 | 6.14 | 6.11 | 3.6 |
  0.8
            Ideal
                       FΙ
                              SI2| 61.4|
                                             57 | 2761 | 5.96 |
                                                                 613.671
```

```
[49]: # Method 2: where df.where(df['cut'] == 'Good').show()
```

```
+----+
|carat| cut|color|clarity|depth|table|price|
| 0.23|Good|
                 Εl
                        VS1 | 56.9
                                      65|
                                            327 | 4.05 | 4.07 | 2.31 |
| 0.31|Good|
                 JΙ
                        SI2| 63.3|
                                      58|
                                            335|4.34|4.35|2.75|
  0.3|Good|
                 JΙ
                        SI1
                                64 l
                                      55 l
                                            339 | 4.25 | 4.28 | 2.73 |
   0.3|Good|
                 JΙ
                        SI1 | 63.4 |
                                      54 l
                                            351|4.23|4.29| 2.7|
  0.3|Good|
                 J|
                        SI1 | 63.8 |
                                      56|
                                            351 | 4.23 | 4.26 | 2.71 |
   0.3 | Good |
                 Ιl
                        SI2 | 63.3 |
                                      56|
                                            351|4.26| 4.3|2.71|
| 0.23|Good|
                                            402|4.06|4.08|2.37|
                 FΙ
                        VS1 | 58.2|
                                      59|
| 0.23|Good|
                 E
                        VS1 64.1
                                      59|
                                            402|3.83|3.85|2.46|
| 0.31|Good|
                 H \mid
                        SI1|
                                64 |
                                      54|
                                            402|4.29|4.31|2.75|
| 0.26|Good|
                 DΙ
                        VS2 65.2
                                      56 l
                                            403|3.99|4.02|2.61|
| 0.26|Good|
                 DI
                        VS1 | 58.4 |
                                      63|
                                            403 | 4.19 | 4.24 | 2.46 |
| 0.32|Good|
                 Ηl
                        SI2 | 63.1 |
                                      56 l
                                            403 | 4.34 | 4.37 | 2.75 |
| 0.32|Good|
                 Η|
                        SI2| 63.8|
                                      561
                                            403|4.36|4.38|2.79|
   0.3|Good|
                 Ιl
                        SI1 | 63.2 |
                                      55 l
                                            405|4.25|4.29| 2.7|
  0.3|Good|
                 Η|
                        SI1 | 63.7 |
                                      57|
                                            554|4.28|4.26|2.72|
```

```
| 0.26|Good|
                   E| VVS1| 57.9| 60| 554|4.22|4.25|2.45|
     | 0.7|Good|
                   Εl
                         VS2| 57.5| 58| 2759|5.85| 5.9|3.38|
                   FΙ
     | 0.7|Good|
                         VS1 | 59.4 | 62 | 2759 | 5.71 | 5.76 | 3.4 |
     | 0.7|Good|
                   Η|
                        VVS2 | 62.1 | 64 | 2767 | 5.62 | 5.65 | 3.5 |
     | 0.71|Good|
                   Εl
                         VS2| 59.2|
                                     61 | 2772 | 5.8 | 5.88 | 3.46 |
     +----+
     only showing top 20 rows
[51]: # Method 2: where
     # select certain columns
     df.where(df['cut'] == 'Good').select('price','cut').show()
     +----+
     |price| cut|
     +----+
     | 327|Good|
       335|Good|
     | 339|Good|
     | 351|Good|
     | 351|Good|
     | 351|Good|
     | 402|Good|
     | 402|Good|
     | 402|Good|
     | 403|Good|
     | 403|Good|
     | 403|Good|
     | 403|Good|
     | 405|Good|
     | 554|Good|
     | 554|Good|
     | 2759|Good|
     | 2759|Good|
     | 2767|Good|
     | 2772|Good|
     +----+
     only showing top 20 rows
[52]: # Unique Values
     # df['cut'].unique()
     df.select("cut").distinct().show()
     +----+
           cutl
     | Premium|
```

```
| Ideal|
| Good|
| Fair|
|Very Good|
```

0.0.3 How to Add Columns & Delete/Drop Columns

- .withColumn()
- .drop()

```
[54]: # Add Columns
     df.withColumn("carat10x",df['carat'] * 10).show()
                                                           уl
     |carat|
                 cut|color|clarity|depth|table|price|
                                                       x
     carat10x|
     --+
     0.23
               Ideal|
                         Εl
                               SI2| 61.5|
                                           55 l
     326|3.95|3.98|2.43|2.3000000000000003|
     | 0.21| Premium|
                                               326|3.89|3.84|2.31|
                         Εl
                               SI1 59.8
                                           61|
     2.11
     1 0.231
                               VS1| 56.9|
                Good
                         Εl
                                           65 l
     327 | 4.05 | 4.07 | 2.31 | 2.3000000000000003 |
     | 0.29| Premium|
                               VS2| 62.4|
                                           58 | 334 | 4.2 | 4.23 | 2.63 |
                         Ιl
     2.9
     | 0.31|
                              SI2| 63.3|
                                               335 | 4.34 | 4.35 | 2.75 |
                Good
                         JΙ
                                           58 l
     3.1
     | 0.24|Very Good|
                         JΙ
                              VVS2| 62.8|
                                           57 | 336 | 3.94 | 3.96 | 2.48 |
     2.4
     | 0.24|Very Good|
                            VVS1| 62.3|
                                           57| 336|3.95|3.98|2.47|
                         Ιl
     2.4
     | 0.26|Very Good|
                              SI1| 61.9|
                                           55 | 337 | 4.07 | 4.11 | 2.53 |
                         Η|
     2.6
     0.22
                Fair|
                         Εl
                               VS2| 65.1|
                                           61 | 337 | 3.87 | 3.78 | 2.49 |
     2.2
     | 0.23|Very Good|
                         Η|
                               VS1| 59.4|
                                           61|
                                               3381
     4|4.05|2.39|2.3000000000000003|
     0.3
                Good
                               SI1|
                                     64|
                                           55| 339|4.25|4.28|2.73|
                         J|
     3.01
                               VS1| 62.8|
     0.23
               Ideal|
                         J|
                                           56| 340|3.93|
     3.9|2.46|2.3000000000000003|
     | 0.22| Premium|
                         FΙ
                               SI1 | 60.4|
                                           61 342 3.88 3.84 2.33
     2.2
     0.31
               Ideal
                         JΙ
                               SI2| 62.2|
                                           54| 344|4.35|4.37|2.71|
```

```
3.1|
| 0.2| Premium|
            E| SI2| 60.2| 62| 345|3.79|3.75|2.27|
2.0|
| 0.32| Premium|
             E| I1| 60.9| 58| 345|4.38|4.42|2.68|
3.2|
0.3
             Ιl
                  SI2| 62| 54| 348|4.31|4.34|2.68|
      | Ideal
3.0|
0.3
            J| SI1| 63.4| 54| 351|4.23|4.29| 2.7|
      Good
3.0|
0.3
      Good J SI1 63.8 56 351 4.23 4.26 2.71
3.0|
| 0.3|Very Good| J| SI1| 62.7| 59| 351|4.21|4.27|2.66|
3.0|
only showing top 20 rows
```

[55]: df.show()

+-	+		+		+	+	+		-+-		+	+
		cut		-	_		_			•	•	z
	0.23				61.5		326					
	0.21	Premium	Εļ	SI1	59.8	61	326	3.8	9 3	.84	12	.31
-	0.23	Good	Εļ	VS1	56.9	65	327	4.0	5 4	.07	12	.31
	0.29	Premium	Ιļ	VS2	62.4	58	334	4.	2 4	.23	12	.63
-	0.31	Good	J	SI2	63.3	58	335	4.3	4 4	.35	12	.75
	0.24	Very Good	J	VVS2	62.8	57	336	3.9	4 3	.96	12	.48
-	0.24	Very Good	Ιļ	VVS1	62.3	57	336	3.9	5 3	.98	12	.47
-	0.26	Very Good	Н	SI1	61.9	55	337	4.0	7 4	.11	12	.53
-	0.22	Fair	Εļ	VS2	65.1	61	337	3.8	7 3	.78	12	.49
-	0.23	Very Good	Н	VS1	59.4	61	338	l	4 4	.05	12	.39
-	0.3	Good	J	SI1	l 64	55	339	4.2	5 4	.28	12	.73
-	0.23	Ideal	J	VS1	62.8	56	340	3.9	3	3.9	12	.46
	0.22	Premium	F	SI1	60.4	61	342	3.8	8 3	.84	12	.33
-	0.31	Ideal	J	SI2	62.2	J 54	344	4.3	5 4	.37	12	.71
-	0.2	Premium	Εļ	SI2	60.2	62	345	3.7	9 3	.75	12	.27
-	0.32	Premium	Εļ	I1	60.9	58	345	4.3	8 4	.42	12	.68
-	0.3	Ideal	Ιļ	SI2	62	J 54	348	4.3	1 4	.34	12	.68
-	0.3	Good	J	SI1	63.4	J 54	351	4.2	3 4	.29	:	2.7
-	0.3	Good	J	SI1	63.8	56	351	4.2	3 4	.26	12	.71
	0.3	Very Good	J	SI1	62.7	59	J 351	4.2	1 4	.27	12	.66
+-	+		+		+	+	+	+	-+-		+	+

```
[63]: df2 = df.withColumn("carat10x",df['carat'] * 10)
```

[65]: df2.show()

++	+	+											
+													
carat cut color clarity depth table price x y z													
carat10x													
++	+	+											
+													
0.23 Ideal E SI2 63	L.5 55												
326 3.95 3.98 2.43 2.30000000000000	0003												
0.21 Premium E SI1 59	9.8 61	326 3.89 3.84 2.31											
2.1													
0.23 Good E VS1 56	5.9 65												
327 4.05 4.07 2.31 2.3000000000000													
0.29 Premium I VS2 62	2.4 58	334 4.2 4.23 2.63											
2.9													
0.31 Good J SI2 63	3.3 58	335 4.34 4.35 2.75											
3.1													
0.24 Very Good	2.8 57	336 3.94 3.96 2.48											
2.4													
0.24 Very Good I VVS1 62	2.3 57	336 3.95 3.98 2.47											
2.4													
0.26 Very Good H SI1 63	1.9 55	337 4.07 4.11 2.53											
2.6													
0.22 Fair E VS2 68	5.1 61	337 3.87 3.78 2.49											
2.2		0001											
0.23 Very Good H VS1 59	9.4 61	338											
4 4.05 2.39 2.3000000000000003	C41 FF1	22014 0514 0010 701											
0.3 Good J SI1	64 55	339 4.25 4.28 2.73											
3.0	ol Eel	24012 021											
0.23 Ideal J VS1 62 3.9 2.46 2.3000000000000003	2.8 56	340 3.93											
0.22 Premium F SI1 60) 41 611	342 3.88 3.84 2.33											
2.2	0.41	342 3.00 3.04 2.33											
0.31 Ideal J SI2 62	0 01 541	3// / 35 / 37 / 71 /											
3.1	2.21 041	011 1.00 1.01 2.11											
0.2 Premium E SI2 60	0.2 62	345 3.79 3.75 2.27											
2.0	021	040 0.70 0.70 2.27											
0.32 Premium E I1 60).9 58	345 4.38 4.42 2.68											
3.2	7.01	010 1.00 1.12 2.00											
0.3 Ideal I SI2	62 54	348 4.31 4.34 2.68											
3.0	021	010 1101 1101 2100											
0.3 Good J SI1 63	3.4 54	351 4.23 4.29 2.7											
3.0	-,,												
0.3 Good J SI1 63													

```
3.0|
| 0.3|Very Good| J| SI1| 62.7| 59| 351|4.21|4.27|2.66|
3.0|
+----+
only showing top 20 rows
```

[67]: # Delete/Drop A Column
df2.drop('carat10x').show()

```
|carat|
             cut|color|clarity|depth|table|price|
                                                            yΙ
0.23
           Ideal
                            SI2| 61.5|
                                          55 l
                                                326|3.95|3.98|2.43|
0.21
         Premium|
                      Εl
                            SI1| 59.8|
                                                326|3.89|3.84|2.31|
                                          61|
0.231
            Good
                      Εl
                            VS1| 56.9|
                                          65 l
                                               327 | 4.05 | 4.07 | 2.31 |
0.291
         Premium|
                      Ιl
                            VS2| 62.4|
                                          581
                                               334 | 4.2 | 4.23 | 2.63 |
0.31
            Good
                      JΙ
                            SI2| 63.3|
                                          58|
                                               335 | 4.34 | 4.35 | 2.75 |
| 0.24|Very Good|
                      JΙ
                           VVS2| 62.8|
                                          57|
                                                336|3.94|3.96|2.48|
| 0.24|Very Good|
                      Ιl
                           VVS1| 62.3|
                                          57|
                                                336|3.95|3.98|2.47|
| 0.26|Very Good|
                      Η|
                            SI1| 61.9|
                                          55|
                                                337 | 4.07 | 4.11 | 2.53 |
0.22
            Fair|
                      Εl
                            VS2| 65.1|
                                          61|
                                                337|3.87|3.78|2.49|
| 0.23|Very Good|
                            VS1| 59.4|
                      НΙ
                                          61 l
                                               3381
                                                       4|4.05|2.39|
0.3
            Good
                      J|
                            SI1|
                                    64 l
                                          55 l
                                               339 | 4.25 | 4.28 | 2.73 |
0.23
           Ideal|
                      J|
                            VS1 | 62.8
                                          56 l
                                               340|3.93| 3.9|2.46|
0.22
         Premium|
                            SI1| 60.4|
                                          61|
                                               342|3.88|3.84|2.33|
                      FΙ
0.31
           Ideal
                      JΙ
                            SI2| 62.2|
                                          54 l
                                               344 | 4.35 | 4.37 | 2.71 |
0.2
                            SI2| 60.2|
                                          62|
         Premium
                                               345 | 3.79 | 3.75 | 2.27 |
1 0.321
         Premium
                      Εl
                             I1 | 60.9 |
                                          58 l
                                               345 | 4.38 | 4.42 | 2.68 |
0.31
           Ideal
                      Ιl
                            SI2|
                                    621
                                          54|
                                               348 | 4.31 | 4.34 | 2.68 |
1 0.31
            Good
                      JΙ
                            SI1| 63.4|
                                          54 l
                                               351 | 4.23 | 4.29 | 2.7 |
0.3
                            SI1| 63.8|
                                          56|
                                               351|4.23|4.26|2.71|
            Good
                      J|
| 0.3|Very Good|
                      JΙ
                            SI1| 62.7|
                                          59 l
                                               351 | 4.21 | 4.27 | 2.66 |
+----+----+----+----+----+
```

only showing top 20 rows

0.0.4 GroupBy

- value counts
- aggregate

```
[69]: # Value Counts
# df['cut'].value_counts()
# df.groupby('cut')[].size()
df.groupBy('cut').count().show()
```

```
cut | count |
     +----+
       Premium | 13791 |
          Ideal | 21551 |
           Good| 4906|
           Fair | 1610 |
     |Very Good|12082|
[70]: # More Groupby
      df.groupBy('price').mean().show()
     +---+
     |price|
     +---+
     | 2904|
     | 3210|
     | 3414|
     36061
     | 3959|
     | 4032|
     | 4821|
     | 4937|
     | 5325|
     | 5645|
     | 5925|
     | 6194|
     | 6240|
     | 6613|
     | 6731|
     | 7273|
     | 7711|
     | 7762|
     1 90091
     9030
     +---+
     only showing top 20 rows
[71]: # Sum of A groupby
      df.groupBy('price').sum().show()
     +---+
     |price|
     +---+
     | 2904|
```

```
3210
| 3414|
1 36061
3959
| 4032|
| 4821|
4937
1 53251
| 5645|
| 5925|
| 6194|
6240
| 6613|
| 6731|
| 7273|
| 7711|
| 7762|
90091
90301
only showing top 20 rows
```

```
[]: # # Aggregation
# df.groupBy('carat').agg('col':'sum')
```

[73]: df.show()

```
cut|color|clarity|depth|table|price|
|carat|
                                                             x
0.23
            Ideal|
                        Εl
                               SI2 | 61.5 |
                                               55|
                                                     326 | 3.95 | 3.98 | 2.43 |
                                                     326|3.89|3.84|2.31|
0.21
          Premium |
                                               61|
                        Εl
                               SI1| 59.8|
0.23
              Goodl
                        Εl
                               VS1 | 56.9 |
                                               65 l
                                                     327 | 4.05 | 4.07 | 2.31 |
0.29
          Premium |
                               VS2| 62.4|
                                               58|
                                                     334 | 4.2 | 4.23 | 2.63 |
0.31
             Goodl
                        JΙ
                               SI2 | 63.3 |
                                               58 l
                                                     335 | 4.34 | 4.35 | 2.75 |
| 0.24|Very Good|
                        J|
                              VVS2| 62.8|
                                               57 l
                                                     336|3.94|3.96|2.48|
| 0.24|Very Good|
                              VVS1| 62.3|
                                                     336|3.95|3.98|2.47|
                        Ιl
                                               57|
| 0.26|Very Good|
                        Η|
                               SI1 | 61.9 |
                                               55 l
                                                     337 | 4.07 | 4.11 | 2.53 |
0.22
                                                     337 | 3.87 | 3.78 | 2.49 |
             Fair|
                        E \mid
                               VS2 | 65.1 |
                                               61|
| 0.23|Very Good|
                               VS1 | 59.4 |
                                                     338|
                                                             4|4.05|2.39|
                        H \mid
                                               61
                                               55 l
  0.3|
              Good |
                        J|
                               SI1|
                                        64 l
                                                     339 | 4.25 | 4.28 | 2.73 |
0.231
                                                     340|3.93| 3.9|2.46|
            Ideal
                        JΙ
                               VS1 | 62.8
                                               56
0.22
          Premium |
                        FΙ
                               SI1 | 60.4 |
                                               61|
                                                     342|3.88|3.84|2.33|
0.31
            Ideal
                        JΙ
                               SI2| 62.2|
                                               54 l
                                                     344 | 4.35 | 4.37 | 2.71 |
   0.21
                               SI2| 60.2|
          Premium |
                        E \mid
                                               621
                                                     345|3.79|3.75|2.27|
0.32
          Premium |
                        Εl
                                I1 | 60.9 |
                                               58 l
                                                     345 | 4.38 | 4.42 | 2.68 |
0.3
            Ideal |
                        Ιl
                                       62|
                                               54|
                                                     348 | 4.31 | 4.34 | 2.68 |
                               SI2|
```

```
0.31
                Good
                             SI1| 63.8|
                                         56 | 351 | 4.23 | 4.26 | 2.71 |
                        JΙ
     | 0.3|Very Good|
                        JΙ
                             SI1| 62.7|
                                         59 l
                                              351 | 4.21 | 4.27 | 2.66 |
    only showing top 20 rows
[74]: df.columns
[74]: ['carat', 'cut', 'color', 'clarity', 'depth', 'table', 'price', 'x', 'y', 'z']
[75]: # Rearrange Columns
     df.select('carat', 'color', 'clarity', 'depth', 'table', 'price', 'x', 'y',
      +----+
     |carat|color|clarity|depth|table|price|
                                           хl
                                                yΙ
    +----+----+-----+
     0.23
                    SI2| 61.5|
                                55| 326|3.95|3.98|2.43|
              Εl
                                                          Ideal
     0.21
                    SI1| 59.8|
              EΙ
                                61 l
                                    326 | 3.89 | 3.84 | 2.31 |
                                                        Premium
     0.23
                    VS1| 56.9|
                                    327 | 4.05 | 4.07 | 2.31 |
                                                           Good
              EΙ
                                65|
     0.29
              Ιl
                    VS2| 62.4|
                                58 l
                                    334 | 4.2 | 4.23 | 2.63 |
                                                        Premium
     0.31
               J|
                    SI2| 63.3|
                                58 | 335 | 4.34 | 4.35 | 2.75 |
                                                           Goodl
     0.24
                   VVS2| 62.8|
                                57| 336|3.94|3.96|2.48|Very Good|
               JΙ
     0.24
                   VVS1| 62.3|
                                57| 336|3.95|3.98|2.47|Very Good|
               Ιl
     0.26
              Η|
                    SI1| 61.9|
                                55| 337|4.07|4.11|2.53|Very Good|
     0.22
                    VS2| 65.1|
                                    337|3.87|3.78|2.49|
              Εl
                                61|
                                                           Fair|
     0.231
              H \mid
                    VS1| 59.4|
                                61|
                                          4|4.05|2.39|Very Good|
                                    338|
                                                           Good
     0.31
               JΙ
                    SI1|
                          64|
                                55 l
                                    339|4.25|4.28|2.73|
     0.231
               JΙ
                    VS1| 62.8|
                                    340|3.93| 3.9|2.46|
                                                          Ideal
                                56|
     1 0.221
              FΙ
                    SI1| 60.4|
                                61 342 3.88 3.84 2.33
                                                        Premium
     0.31
                    SI21 62.21
                                54 344 4.35 4.37 2.71
                                                          Ideal
               JΙ
     0.2
              Εl
                    SI2| 60.2|
                                62| 345|3.79|3.75|2.27| Premium|
     0.32
              EΙ
                    I1| 60.9|
                                58 | 345 | 4.38 | 4.42 | 2.68 |
                                                        Premium |
     0.3
              Ιl
                    SI2|
                          62|
                                54 348 4.31 4.34 2.68
                                                          Ideal
     0.3
               J|
                    SI1 | 63.4 |
                                54 | 351 | 4.23 | 4.29 | 2.7 |
                                                           Good
     0.31
                                56 | 351 | 4.23 | 4.26 | 2.71 |
               JΙ
                    SI1| 63.8|
                                                           Good
     0.3
                    SI1| 62.7|
                                59| 351|4.21|4.27|2.66|Very Good|
               J|
     +----+---+-----+-----+
    only showing top 20 rows
[76]: # Assign DF to a New DF
     new_df = df.select('carat', 'color', 'clarity', 'depth', 'table', 'price', 'x', |
      [77]: new df
```

1 0.31

Good

JΙ

SI1| 63.4|

54 l

351 4.23 4.29 | 2.7 |

```
[77]: DataFrame[carat: string, color: string, clarity: string, depth: string, table:
      string, price: string, x: string, y: string, z: string, cut: string]
[78]: # Check Datatype
      new_df.dtypes
[78]: [('carat', 'string'),
       ('color', 'string'),
       ('clarity', 'string'),
       ('depth', 'string'),
       ('table', 'string'),
       ('price', 'string'),
       ('x', 'string'),
       ('y', 'string'),
       ('z', 'string'),
       ('cut', 'string')]
[80]: # Check For the Schema
      df.printSchema()
     root
      |-- carat: string (nullable = true)
      |-- cut: string (nullable = true)
      |-- color: string (nullable = true)
      |-- clarity: string (nullable = true)
      |-- depth: string (nullable = true)
      |-- table: string (nullable = true)
      |-- price: string (nullable = true)
      |-- x: string (nullable = true)
      |-- y: string (nullable = true)
      |-- z: string (nullable = true)
[81]: # Check type of DF
      type(df)
[81]: pyspark.sql.dataframe.DataFrame
     0.0.5 Saving DataFrames as CSV, parquet etc
[83]: # Save
      new_df.write.format('csv').option('header','true').save("diamond_clean.csv")
[84]: !ls
      diamond_clean.csv
                          diamonds.csv 'PySpark Crash Course.ipynb'
                                                                        work
```

```
[]:  # Save as parquet new_df.write.format('parquet').save("diamond_clean.parquet")
```

```
0.0.6 Making SQL Queries
         • parse in the spark.SparkContext
         • sqlContext
[85]: from pyspark.sql import SQLContext
[92]: dir(spark)
[92]: ['Builder',
        '__class__',
        '__delattr__',
        '__dict__',
        '__dir__',
        '__doc__',
        '__enter__',
        '__eq__',
'__exit__',
        '__format__',
       '__ge__',
'__getattribute__',
        '__gt__',
        '__hash__',
       '__init__',
'__init_subclass__',
        '__le__',
        '__lt__',
        '__module__',
        '__ne__',
        '__new__',
        '__reduce__',
'__reduce_ex__',
        '__repr__',
        '__setattr__',
        '__sizeof__',
        '__str__',
         __subclasshook__',
        '__weakref__',
        '_activeSession',
        '_convert_from_pandas',
        '_createFromLocal',
```

'_createFromRDD',
'_create_dataframe',

'_create_from_pandas_with_arrow',

```
'_get_numpy_record_dtype',
      '_inferSchema',
      '_inferSchemaFromList',
      '_instantiatedSession',
      '_jsc',
      '_jsparkSession',
      '_jvm',
      '_jwrapped',
      '_repr_html_',
      '_sc',
      '_wrapped',
      'builder',
      'catalog',
      'conf',
      'createDataFrame',
      'getActiveSession',
      'newSession',
      'range',
      'read',
      'readStream',
      'sparkContext',
      'sql',
      'stop',
      'streams',
      'table',
      'udf',
      'version']
[93]: # Create A Spark Context From the Spark Session
     sc = spark.sparkContext
[94]: # Parse into the SQLContext
     sqlContext = SQLContext(sc)
[96]: # Register Current DataFrame As Temporal Table
     df.registerTempTable("DiamondsTable")
[97]: # Making QUeries
     sqlContext.sql('SELECT * FROM DiamondsTable').show()
                      ---+----+
                cut|color|clarity|depth|table|price|
    +----+
     1 0.231
              Ideall
                        Εl
                             SI2| 61.5|
                                         55| 326|3.95|3.98|2.43|
     | 0.21| Premium|
                       Εl
                             SI1| 59.8|
                                         61| 326|3.89|3.84|2.31|
     1 0.231
               Goodl
                       Εl
                             VS1| 56.9|
                                         65 | 327 | 4.05 | 4.07 | 2.31 |
```

'_create_shell_session',

```
| 0.29| Premium|
                      Ιl
                            VS2| 62.4|
                                          58|
                                               334 | 4.2 | 4.23 | 2.63 |
| 0.31|
            Good
                            SI2| 63.3|
                                               335|4.34|4.35|2.75|
                      JΙ
                                          581
| 0.24|Very Good|
                      JΙ
                           VVS2| 62.8|
                                          57|
                                               336|3.94|3.96|2.48|
| 0.24|Very Good|
                      Ιl
                           VVS1| 62.3|
                                          57|
                                               336|3.95|3.98|2.47|
| 0.26|Very Good|
                            SI1| 61.9|
                                               337 | 4.07 | 4.11 | 2.53 |
                      НΙ
                                          55 l
0.22
            Fair|
                      Εl
                            VS2 | 65.1|
                                          61|
                                               337|3.87|3.78|2.49|
| 0.23|Very Good|
                      Η|
                            VS1| 59.4|
                                          61|
                                               338
                                                      4|4.05|2.39|
0.3
            Good
                      J|
                            SI1
                                   64|
                                          55 l
                                               339 | 4.25 | 4.28 | 2.73 |
0.23
           Ideal|
                            VS1 | 62.8|
                                               340|3.93| 3.9|2.46|
                      J|
                                          56|
0.22
        Premium
                      FΙ
                            SI1| 60.4|
                                          61|
                                               342|3.88|3.84|2.33|
0.31
                            SI2| 62.2|
                                               344 | 4.35 | 4.37 | 2.71 |
           Ideal|
                      J|
                                          54|
0.2
        Premium|
                                          621
                                               345|3.79|3.75|2.27|
                      E
                            SI2| 60.2|
0.32
        Premium|
                      E \mid
                             I1| 60.9|
                                          581
                                               345 | 4.38 | 4.42 | 2.68 |
0.31
           Ideal
                                   621
                                               348 | 4.31 | 4.34 | 2.68 |
                      Ιl
                            SI2|
                                          54|
0.31
                                               351 | 4.23 | 4.29 | 2.7 |
            Good
                      JΙ
                            SI1| 63.4|
                                          54 l
0.31
            Good
                            SI1| 63.8|
                                          561
                                               351 | 4.23 | 4.26 | 2.71 |
                      JΙ
| 0.3|Very Good|
                      J|
                            SI1| 62.7|
                                          59|
                                               351 | 4.21 | 4.27 | 2.66 |
+----+
                         ----+---+
```

```
[98]: # Can also use it to work with DataFrames
dir(sqlContext)
```

```
[98]: ['__class__',
       '__delattr__',
       '__dict__',
       '__dir__',
       '__doc__',
         __eq__',
       '__format__',
       '__ge__',
       '__getattribute__',
       '__gt__',
       '__hash__',
       '__init__',
       '__init_subclass__',
       '__le__',
       '__lt__',
       '__module__',
       '__ne__',
       '__new__',
       '__reduce__',
       '__reduce_ex__',
       '__repr__',
       '__setattr__',
       '__sizeof__',
```

```
'__str__',
       '__subclasshook__',
       '__weakref__',
       '_conf',
       '_inferSchema',
       '_instantiatedContext',
       '_jsc',
       '_jsqlContext',
       '_jvm',
       '_sc',
       '_ssql_ctx',
       'cacheTable',
       'clearCache',
       'createDataFrame',
       'createExternalTable',
       'dropTempTable',
       'getConf',
       'getOrCreate',
       'newSession',
       'range',
       'read',
       'readStream',
       'registerDataFrameAsTable',
       'registerFunction',
       'registerJavaFunction',
       'setConf',
       'sparkSession',
       'sql',
       'streams',
       'table',
       'tableNames',
       'tables',
       'udf',
       'uncacheTable']
[99]: # Thanks For Watching
      # Jesus Saves @JCharisTech
      # Jesse E.Agbe(JCharis)
      # 2021
 []:
```