Initialize Spark

In [1]:
```
import findspark
```

In [2]:
```
findspark.init('/home/ubuntu/spark-2.4.5-bin-hadoop2.7')
```

In [3]:
```
import pyspark
```

In [4]:
```
from pyspark.sql import SparkSession
```

In [5]:
```
spark = SparkSession.builder.appName('Basics').getOrCreate()
```

Read data & change schema

In [6]:
```
df = spark.read.csv('udemy_courses.csv', header=True)
```

In [7]:
```
df.show()
```

```
+---------+--------------------+--------------------+-------+-----+-----
---------+----------+------------+------------------+----------------+
--------------------+---------------+
|course_id|        course_title|                 url|is_paid|price|num_s
ubscribers|num_reviews|num_lectures|             level|content_duration|
published_timestamp|         subject|
+---------+--------------------+--------------------+-------+-----+-----
---------+----------+------------+------------------+----------------+
--------------------+---------------+
|  1070968|Ultimate Investme...|https://www.udemy...|   True|  200|
2147|         23|          51|        All Levels|       1.5 hours|2017-0
1-18T20:58:58Z|Business Finance|
|  1113822|Complete GST Cour...|https://www.udemy...|   True|   75|
2792|        923|         274|        All Levels|        39 hours|2017-0
3-09T16:34:20Z|Business Finance|
|  1006314|Financial Modelin...|https://www.udemy...|   True|   45|
2174|         74|          51|Intermediate Level|       2.5 hours|2016-1
2-19T19:26:30Z|Business Finance|
|  1210588|Beginner to Pro -...|https://www.udemy...|   True|   95|
2451|         11|          36|        All Levels|         3 hours|2017-0
5-30T20:07:24Z|Business Finance|
|  1011058|How To Maximize Y...|https://www.udemy...|   True|  200|
1276|         45|          26|Intermediate Level|         2 hours|2016-1
2-13T14:57:18Z|Business Finance|
|   192870|Trading Penny Sto...|https://www.udemy...|   True|  150|
9221|        138|          25|        All Levels|         3 hours|2014-0
5-02T15:13:30Z|Business Finance|
|   739964|Investing And Tra...|https://www.udemy...|   True|   65|
1540|        178|          26|    Beginner Level|          1 hour|2016-0
2-21T18:23:12Z|Business Finance|
|   403100|Trading Stock Cha...|https://www.udemy...|   True|   95|
2917|        148|          23|        All Levels|       2.5 hours|2015-0
1-30T22:13:03Z|Business Finance|
|   476268|Options Trading 3...|https://www.udemy...|   True|  195|
5172|         34|          38|      Expert Level|       2.5 hours|2015-0
5-28T00:14:03Z|Business Finance|
|  1167710|The Only Investme...|https://www.udemy...|   True|  200|
827|         14|          15|        All Levels|         1 hour|2017-04
-18T18:13:32Z|Business Finance|
|   592338|Forex Trading Sec...|https://www.udemy...|   True|  200|
4284|         93|          76|        All Levels|         5 hours|2015-0
9-11T16:47:02Z|Business Finance|
|   975046|Trading Options W...|https://www.udemy...|   True|  200|
1380|         42|          17|        All Levels|          1 hour|2016-1
0-18T22:52:31Z|Business Finance|
|   742602|Financial Managem...|https://www.udemy...|   True|   30|
3607|         21|          19|        All Levels|       1.5 hours|2016-0
2-03T18:04:01Z|Business Finance|
|   794151|Forex Trading Cou...|https://www.udemy...|   True|  195|
4061|         52|          16|        All Levels|         2 hours|2016-0
3-16T15:40:19Z|Business Finance|
|  1196544|Python Algo Tradi...|https://www.udemy...|   True|  200|
294|         19|          42|        All Levels|         7 hours|2017-04
-28T16:41:44Z|Business Finance|
```

```
|   504036|Short Selling: Le...|https://www.udemy...|   True|   75|
2276|        106|          19|Intermediate Level|       1.5 hours|2015-0
6-22T21:18:35Z|Business Finance|
|   719698|Basic Technical A...|https://www.udemy...|   True|   20|
4919|         79|          16|    Beginner Level|       1.5 hours|2016-0
1-08T17:21:26Z|Business Finance|
|   564966|The Complete Char...|https://www.udemy...|   True|  200|
2666|        115|          52|        All Levels|         4 hours|2015-0
8-10T21:07:35Z|Business Finance|
|   606928|7 Deadly Mistakes...|https://www.udemy...|   True|   50|
5354|         24|          23|        All Levels|       1.5 hours|2015-0
9-21T18:10:34Z|Business Finance|
|    58977|Financial Stateme...|https://www.udemy...|   True|   95|
8095|        249|          12|    Beginner Level|         35 mins|2013-0
6-09T00:21:26Z|Business Finance|
+---------+--------------------+--------------------+-------+-----+-----
----------+-----------+------------+------------------+----------------+
--------------------+----------------+
only showing top 20 rows
```

In [8]:
```
df.printSchema()
```

```
root
 |-- course_id: string (nullable = true)
 |-- course_title: string (nullable = true)
 |-- url: string (nullable = true)
 |-- is_paid: string (nullable = true)
 |-- price: string (nullable = true)
 |-- num_subscribers: string (nullable = true)
 |-- num_reviews: string (nullable = true)
 |-- num_lectures: string (nullable = true)
 |-- level: string (nullable = true)
 |-- content_duration: string (nullable = true)
 |-- published_timestamp: string (nullable = true)
 |-- subject: string (nullable = true)
```

In [9]:
```
from pyspark.sql.types import IntegerType, StringType, DateType, TimestampTyp
```

In [10]:
```
field = [StructField('course_id', StringType(), True),
         StructField('course_title', StringType(), True),
         StructField('url', StringType(), True),
         StructField('is_paid', BooleanType(), True),
         StructField('price', IntegerType(), True),
         StructField('num_subscribers', IntegerType(), True),
         StructField('num_reviews', IntegerType(), True),
         StructField('num_lectures', IntegerType(), True),
         StructField('level', StringType(), True),
         StructField('content_duration', StringType(), True),
         StructField('published_timestamp', TimestampType(), True),
         StructField('subject', StringType(), True)
        ]
```

In [11]:
```
schema = StructType(fields=field)
```

In [12]:
```
df = spark.read.csv('udemy_courses.csv', schema=schema, header=True)
```

In [13]:
```
df.printSchema()
```

```
root
 |-- course_id: string (nullable = true)
 |-- course_title: string (nullable = true)
 |-- url: string (nullable = true)
 |-- is_paid: boolean (nullable = true)
 |-- price: integer (nullable = true)
 |-- num_subscribers: integer (nullable = true)
 |-- num_reviews: integer (nullable = true)
 |-- num_lectures: integer (nullable = true)
 |-- level: string (nullable = true)
 |-- content_duration: string (nullable = true)
 |-- published_timestamp: timestamp (nullable = true)
 |-- subject: string (nullable = true)
```

In [14]:
```
df.describe('price').show()
```

```
+-------+-----------------+
|summary|            price|
+-------+-----------------+
|  count|             3378|
|   mean|72.07815275310834|
| stddev|60.20277493162526|
|    min|               20|
|    max|              200|
+-------+-----------------+
```

In [15]:
```
df.columns
```

Out[15]:
```
['course_id',
 'course_title',
 'url',
 'is_paid',
 'price',
 'num_subscribers',
 'num_reviews',
 'num_lectures',
 'level',
 'content_duration',
 'published_timestamp',
 'subject']
```

In [16]:
```
df.describe('num_subscribers').show()
```

```
+-------+-----------------+
|summary|  num_subscribers|
+-------+-----------------+
|  count|             3689|
|   mean|3189.809704526972|
| stddev|9490.950565928886|
|    min|                0|
|    max|           268923|
+-------+-----------------+
```

DataFrame manipulation

In [17]:
```
df.show(3, truncate=True)
```

```
+---------+--------------------+--------------------+-------+-----+--------
-------+----------+------------+------------------+----------------+------
-------------+----------------+
|course_id|        course_title|                 url|is_paid|price|num_subs
cribers|num_reviews|num_lectures|             level|content_duration|publis
hed_timestamp|         subject|
+---------+--------------------+--------------------+-------+-----+--------
-------+----------+------------+------------------+----------------+------
-------------+----------------+
|  1070968|Ultimate Investme...|https://www.udemy...|   true|  200|
2147|         23|          51|        All Levels|       1.5 hours|2017-01-1
8 20:58:58|Business Finance|
|  1113822|Complete GST Cour...|https://www.udemy...|   true|   75|
2792|        923|         274|        All Levels|        39 hours|2017-03-0
9 16:34:20|Business Finance|
|  1006314|Financial Modelin...|https://www.udemy...|   true|   45|
2174|         74|          51|Intermediate Level|       2.5 hours|2016-12-1
9 19:26:30|Business Finance|
+---------+--------------------+--------------------+-------+-----+--------
-------+----------+------------+------------------+----------------+------
-------------+----------------+
only showing top 3 rows
```

In [18]:
```
df.count(), len(df.columns)
```

Out[18]: (3689, 12)

In [19]:
```
df.columns
```

Out[19]:
```
['course_id',
 'course_title',
 'url',
 'is_paid',
 'price',
 'num_subscribers',
 'num_reviews',
 'num_lectures',
 'level',
 'content_duration',
 'published_timestamp',
 'subject']
```

In [20]:
```
df.describe().show()
```

```
+-------+-----------------+---------------------------------+-------------
------+-----------------+-----------------+------------------+-------------
-----+------------------+----------------+----------------+
|summary|        course_id|                     course_title|
url|            price|  num_subscribers|       num_reviews|      num_lectur
es|             level|content_duration|         subject|
+-------+-----------------+---------------------------------+-------------
------+-----------------+-----------------+------------------+-------------
-----+------------------+----------------+----------------+
|  count|             3378|                             3378|
3378|             3378|             3378|              3378|             3
378|              3378|            3378|            3372|
|   mean|674494.0941385435|                             null|
null|72.07815275310834|2425.000888099467|131.08466548253404| 41.77205447010
065|              null|             0.0|            null|
| stddev|341059.9160009458|                             null|
null|60.20277493162526|6351.603756193211| 924.7177555270208|51.971013315449
625|              null|             NaN|            null|
|    min|          1000010|             #1 Piano Hand Coo...|https://www.ud
emy...|               20|                0|                 0|
0|        All Levels|              0|Business Finance|
|    max|            99986|７日でマスター　ビギナー向け A...|https://www.udem
y...|              200|           121584|             27445|
779|Intermediate Level|       9.5 hours| Web Development|
+-------+-----------------+---------------------------------+-------------
------+-----------------+-----------------+------------------+-------------
-----+------------------+----------------+----------------+
```

In [21]:
```
df.describe('price').show()
```

```
+-------+-----------------+
|summary|            price|
+-------+-----------------+
|  count|             3378|
|   mean|72.07815275310834|
| stddev|60.20277493162526|
|    min|               20|
|    max|              200|
+-------+-----------------+
```

In [22]:
```
df.describe('num_subscribers', 'num_reviews', 'num_lectures').show()
```

```
+-------+-----------------+------------------+------------------+
|summary|  num_subscribers|       num_reviews|       num_lectures|
+-------+-----------------+------------------+------------------+
|  count|             3689|              3689|              3689|
|   mean|3189.809704526972|156.22255353754406| 40.05801030089455|
| stddev|9490.950565928886| 934.3341656220671|50.331707687614994|
|    min|                0|                 0|                 0|
|    max|           268923|             27445|               779|
+-------+-----------------+------------------+------------------+
```

In [23]:
```
df.select('price').show()
```

```
+-----+
|price|
+-----+
|  200|
|   75|
|   45|
|   95|
|  200|
|  150|
|   65|
|   95|
|  195|
|  200|
|  200|
|  200|
|   30|
|  195|
|  200|
|   75|
|   20|
|  200|
|   50|
|   95|
+-----+
only showing top 20 rows
```

In [24]:
```
df.select(['level','price']).show()
```

```
+------------------+-----+
|             level|price|
+------------------+-----+
|        All Levels|  200|
|        All Levels|   75|
|Intermediate Level|   45|
|        All Levels|   95|
|Intermediate Level|  200|
|        All Levels|  150|
|    Beginner Level|   65|
|        All Levels|   95|
|      Expert Level|  195|
|        All Levels|  200|
|        All Levels|  200|
|        All Levels|  200|
|        All Levels|   30|
|        All Levels|  195|
|        All Levels|  200|
|Intermediate Level|   75|
|    Beginner Level|   20|
|        All Levels|  200|
|        All Levels|   50|
|    Beginner Level|   95|
+------------------+-----+
only showing top 20 rows
```

In [25]:
```
df.distinct().count()
```

Out[25]: 3373

In [26]:
```
df.select('level').distinct().count()
```

Out[26]: 5

In [27]:
```
df.crosstab('level', 'is_paid').show()
```

```
+------------------+-----+----+----+
|     level_is_paid|false|null|true|
+------------------+-----+----+----+
|              null|    0|   1|   0|
|      Expert Level|    0|   0|  58|
|Intermediate Level|   30|   0| 392|
|    Beginner Level|  158|   0|1116|
|        All Levels|  122|   0|1812|
+------------------+-----+----+----+
```

In [28]:
```
df.select('level', 'content_duration').dropDuplicates().show()
```

```
+------------------+----------------+
|             level|content_duration|
+------------------+----------------+
|    Beginner Level|       2.5 hours|
|        All Levels|         36 mins|
|        All Levels|      22.5 hours|
|    Beginner Level|       5.5 hours|
|Intermediate Level|         35 mins|
|    Beginner Level|      20.5 hours|
|Intermediate Level|         30 mins|
|        All Levels|        51 hours|
|        All Levels|       9.5 hours|
|        All Levels|      20.5 hours|
|        All Levels|         8 hours|
|        All Levels|        21 hours|
|      Expert Level|       3.5 hours|
|    Beginner Level|        15 hours|
|    Beginner Level|      44.5 hours|
|        All Levels|      46.5 hours|
|    Beginner Level|         34 mins|
|        All Levels|        20 hours|
|    Beginner Level|         40 mins|
|    Beginner Level|         2 hours|
+------------------+----------------+
only showing top 20 rows
```

In [29]:
```
df.filter((df['price']>70) & (df['num_subscribers']>10000)).show()
```

```
+---------+--------------------+--------------------+-------+-----+-----
---------+-----------+------------+--------------+----------------+----
--------------+-------------------+
|course_id|        course_title|                 url|is_paid|price|num_s
ubscribers|num_reviews|num_lectures|         level|content_duration|publ
ished_timestamp|            subject|
+---------+--------------------+--------------------+-------+-----+-----
---------+-----------+------------+--------------+----------------+----
--------------+-------------------+
|   506568|Create A Business...|https://www.udemy...|   true|   75|
10149|         83|          16|    All Levels|         2 hours|2015-05-2
6 17:25:46|   Business Finance|
|   308690|Forex Trading A-Z...|https://www.udemy...|   true|  195|
16900|       2476|          52|Beginner Level|       5.5 hours|2014-12-1
2 23:58:39|   Business Finance|
|   888716|Introduction to F...|https://www.udemy...|   true|  200|
11441|       1118|          61|    All Levels|       4.5 hours|2016-06-2
8 06:12:23|   Business Finance|
|   321410|Beginner to Pro i...|https://www.udemy...|   true|  195|
22257|       2697|         138|    All Levels|       7.5 hours|2014-11-2
5 23:00:40|   Business Finance|
|   648826|The Complete Fina...|https://www.udemy...|   true|  195|
24481|       2347|         174|    All Levels|        10 hours|2016-01-2
1 01:38:48|   Business Finance|
|   301442|Black Algo Tradin...|https://www.udemy...|   true|  200|
20195|       1113|         227|    All Levels|        16 hours|2014-10-2
7 22:01:36|   Business Finance|
|   640100|Accounting & Fina...|https://www.udemy...|   true|  150|
10042|        594|          43|    All Levels|         3 hours|2015-10-2
2 00:03:48|   Business Finance|
|   260470|Forex Robots: Exp...|https://www.udemy...|   true|  200|
10603|        872|          42|    All Levels|       5.5 hours|2014-10-1
2 21:19:11|   Business Finance|
|   325834|Learn to Trade fo...|https://www.udemy...|   true|   95|
10605|         71|          77|    All Levels|         3 hours|2014-10-2
4 18:13:49|   Business Finance|
|    43319|Options Trading B...|https://www.udemy...|   true|  180|
10100|        985|          45|Beginner Level|        11 hours|2013-02-2
5 11:36:06|   Business Finance|
|   401784|Options Trading I...|https://www.udemy...|   true|   95|
12394|        218|          30|Beginner Level|       2.5 hours|2015-02-2
0 21:39:41|   Business Finance|
|   116128|CPA 101: How To M...|https://www.udemy...|   true|  100|
11517|         92|          21|    All Levels|       1.5 hours|2013-11-0
9 21:46:52|   Business Finance|
|  1120554|Canva Graphics De...|https://www.udemy...|   true|  200|
12340|        124|          46|    All Levels|       2.5 hours|2017-03-0
7 03:54:27|     Graphic Design|
|   749542|Graphic Design Bo...|https://www.udemy...|   true|  200|
15276|       1740|          65|Beginner Level|         8 hours|2016-05-1
3 00:03:03|     Graphic Design|
|   874012|The Ultimate Draw...|https://www.udemy...|   true|  150|
26742|       2379|          62|Beginner Level|        11 hours|2017-01-2
3 00:20:05|     Graphic Design|
```

```
|   820194|Photoshop for Ent...|https://www.udemy...|   true|  200|
36288|        737|          63|    All Levels|         5 hours|2016-06-0
9 01:57:03|     Graphic Design|
|    62721|Anatomy for Figur...|https://www.udemy...|   true|   95|
15500|        754|          65|    All Levels|      68.5 hours|2013-10-1
6 11:37:30|     Graphic Design|
|   897238|Canva Graphics De...|https://www.udemy...|   true|  200|
18303|        202|          54|    All Levels|       3.5 hours|2016-07-2
3 00:41:07|     Graphic Design|
|   178044|How To Make Graph...|https://www.udemy...|   true|  200|
24857|         35|          11|    All Levels|       1.5 hours|2014-03-1
5 21:53:19|     Graphic Design|
|   238934|Pianoforall - Inc...|https://www.udemy...|   true|  200|
75499|       7676|         362|    All Levels|        30 hours|2014-08-0
7 06:27:51|Musical Instruments|
+---------+--------------------+--------------------+-------+-----+-----
----------+-----------+------------+-------------+----------------+----
---------------+-------------------+
only showing top 20 rows
```

In [30]:
```
df.groupby('level').agg({'price':'mean'}).show()
```

```
+------------------+-----------------+
|             level|       avg(price)|
+------------------+-----------------+
|      Expert Level|91.12068965517241|
|Intermediate Level| 66.7984693877551|
|              null|             null|
|        All Levels|77.90562913907284|
|    Beginner Level|63.48118279569893|
+------------------+-----------------+
```

In [31]:
```
df.groupby('level').mean().show()
```

```
+------------------+-----------------+--------------------+----------------
--+------------------+
|             level|       avg(price)|avg(num_subscribers)|  avg(num_review
s)| avg(num_lectures)|
+------------------+-----------------+--------------------+----------------
--+------------------+
|      Expert Level|91.12068965517241|    865.448275862069|40.2241379310344
84|30.775862068965516|
|Intermediate Level| 66.7984693877551|   1375.204081632653| 82.448979591836
73|37.329081632653065|
|              null|             null|                null|              nu
ll|              null|
|        All Levels|77.90562913907284|  2919.7041942604856|189.721302428256
08| 47.33664459161148|
|    Beginner Level|63.48118279569893|   2071.570788530466| 57.684587813620
07|34.869175627240146|
+------------------+-----------------+--------------------+----------------
--+------------------+
```

In [32]:
```
df.groupby('level').count().show()
```

```
+------------------+-----+
|             level|count|
+------------------+-----+
|      Expert Level|   58|
|Intermediate Level|  422|
|                52|    1|
|        All Levels| 1934|
|    Beginner Level| 1274|
+------------------+-----+
```

In [33]:
```
df.select('level').distinct().show()
```

```
+------------------+
|             level|
+------------------+
|      Expert Level|
|Intermediate Level|
|                52|
|        All Levels|
|    Beginner Level|
+------------------+
```

In [34]:
```
seventyfour = df.filter(df['price']==75).collect()
```

In [35]:
```
from pyspark.sql.functions import countDistinct, least, stddev, variance, mea
```

In [36]:
```
from pyspark.sql.functions import format_number
```

In [37]:
```
price_avg = df.select(mean('price'))
```

In [38]:
```
price_avg.show()
```

```
+-----------------+
|       avg(price)|
+-----------------+
|72.07815275310834|
+-----------------+
```

In [39]:
```
price_avg.select(format_number('avg(price)',2).alias('average_price')).show()
```

```
+-------------+
|average_price|
+-------------+
|        72.08|
+-------------+
```

Missing Data

In [40]:
```
df_trial = df.na.drop()
```

In [41]:
```
df_trial.count()
```

Out[41]: 3372

In [42]:
```
df.count()
```

Out[42]: 3689

In [43]:
```
df.na.fill(9999)
```

Out[43]:
```
DataFrame[course_id: string, course_title: string, url: string, is_paid: bo
olean, price: int, num_subscribers: int, num_reviews: int, num_lectures: in
t, level: string, content_duration: string, published_timestamp: timestamp,
subject: string]
```

In [44]:
```
df.columns
```

Out[44]:
```
['course_id',
 'course_title',
 'url',
 'is_paid',
 'price',
 'num_subscribers',
 'num_reviews',
 'num_lectures',
 'level',
 'content_duration',
 'published_timestamp',
 'subject']
```

DateTime Pyspark

In [45]:
```
from pyspark.sql.functions import format_number, date_format, dayofyear, dayo
```

In [46]:
```
df.show()
```

```
+---------+--------------------+--------------------+-------+-----+--------
-------+----------+-----------+-----------------+---------------+------
------------+---------------+
|course_id|        course_title|                 url|is_paid|price|num_subs
cribers|num_reviews|num_lectures|            level|content_duration|publis
hed_timestamp|         subject|
+---------+--------------------+--------------------+-------+-----+--------
-------+----------+-----------+-----------------+---------------+------
------------+---------------+
|  1070968|Ultimate Investme...|https://www.udemy...|   true|  200|
2147|         23|          51|        All Levels|       1.5 hours|2017-01-1
8 20:58:58|Business Finance|
|  1113822|Complete GST Cour...|https://www.udemy...|   true|   75|
2792|        923|         274|        All Levels|        39 hours|2017-03-0
9 16:34:20|Business Finance|
|  1006314|Financial Modelin...|https://www.udemy...|   true|   45|
2174|         74|          51|Intermediate Level|       2.5 hours|2016-12-1
9 19:26:30|Business Finance|
|  1210588|Beginner to Pro -...|https://www.udemy...|   true|   95|
2451|         11|          36|        All Levels|         3 hours|2017-05-3
0 20:07:24|Business Finance|
|  1011058|How To Maximize Y...|https://www.udemy...|   true|  200|
1276|         45|          26|Intermediate Level|         2 hours|2016-12-1
3 14:57:18|Business Finance|
|   192870|Trading Penny Sto...|https://www.udemy...|   true|  150|
9221|        138|          25|        All Levels|         3 hours|2014-05-0
2 15:13:30|Business Finance|
|   739964|Investing And Tra...|https://www.udemy...|   true|   65|
1540|        178|          26|    Beginner Level|          1 hour|2016-02-2
1 18:23:12|Business Finance|
|   403100|Trading Stock Cha...|https://www.udemy...|   true|   95|
2917|        148|          23|        All Levels|       2.5 hours|2015-01-3
0 22:13:03|Business Finance|
|   476268|Options Trading 3...|https://www.udemy...|   true|  195|
5172|         34|          38|      Expert Level|       2.5 hours|2015-05-2
8 00:14:03|Business Finance|
|  1167710|The Only Investme...|https://www.udemy...|   true|  200|
827|         14|          15|        All Levels|         1 hour|2017-04-18
18:13:32|Business Finance|
|   592338|Forex Trading Sec...|https://www.udemy...|   true|  200|
4284|         93|          76|        All Levels|         5 hours|2015-09-1
1 16:47:02|Business Finance|
|   975046|Trading Options W...|https://www.udemy...|   true|  200|
1380|         42|          17|        All Levels|          1 hour|2016-10-1
8 22:52:31|Business Finance|
|   742602|Financial Managem...|https://www.udemy...|   true|   30|
3607|         21|          19|        All Levels|       1.5 hours|2016-02-0
3 18:04:01|Business Finance|
|   794151|Forex Trading Cou...|https://www.udemy...|   true|  195|
4061|         52|          16|        All Levels|         2 hours|2016-03-1
6 15:40:19|Business Finance|
|  1196544|Python Algo Tradi...|https://www.udemy...|   true|  200|
294|         19|          42|        All Levels|         7 hours|2017-04-28
16:41:44|Business Finance|
```

```
|   504036|Short Selling: Le...|https://www.udemy...|   true|   75|
2276|        106|          19|Intermediate Level|       1.5 hours|2015-06-2
2 21:18:35|Business Finance|
|   719698|Basic Technical A...|https://www.udemy...|   true|   20|
4919|         79|          16|    Beginner Level|       1.5 hours|2016-01-0
8 17:21:26|Business Finance|
|   564966|The Complete Char...|https://www.udemy...|   true|  200|
2666|        115|          52|        All Levels|         4 hours|2015-08-1
0 21:07:35|Business Finance|
|   606928|7 Deadly Mistakes...|https://www.udemy...|   true|   50|
5354|         24|          23|        All Levels|       1.5 hours|2015-09-2
1 18:10:34|Business Finance|
|    58977|Financial Stateme...|https://www.udemy...|   true|   95|
8095|        249|          12|    Beginner Level|         35 mins|2013-06-0
9 00:21:26|Business Finance|
+---------+--------------------+--------------------+-------+-----+--------
-------+-----------+------------+------------------+----------------+------
-------------+----------------+
only showing top 20 rows
```

In [47]:
```
df_date = df.withColumn("year", year(df['published_timestamp']))
```

In [48]:
```
df_date.show()
```

```
+---------+--------------------+--------------------+-------+-----+--------
-------+----------+-----------+-----------------+---------------+------
------------+---------------+----+
|course_id|        course_title|                 url|is_paid|price|num_subs
cribers|num_reviews|num_lectures|            level|content_duration|publis
hed_timestamp|         subject|year|
+---------+--------------------+--------------------+-------+-----+--------
-------+----------+-----------+-----------------+---------------+------
------------+---------------+----+
|  1070968|Ultimate Investme...|https://www.udemy...|   true|  200|
2147|         23|          51|        All Levels|       1.5 hours|2017-01-1
8 20:58:58|Business Finance|2017|
|  1113822|Complete GST Cour...|https://www.udemy...|   true|   75|
2792|        923|         274|        All Levels|        39 hours|2017-03-0
9 16:34:20|Business Finance|2017|
|  1006314|Financial Modelin...|https://www.udemy...|   true|   45|
2174|         74|          51|Intermediate Level|       2.5 hours|2016-12-1
9 19:26:30|Business Finance|2016|
|  1210588|Beginner to Pro -...|https://www.udemy...|   true|   95|
2451|         11|          36|        All Levels|         3 hours|2017-05-3
0 20:07:24|Business Finance|2017|
|  1011058|How To Maximize Y...|https://www.udemy...|   true|  200|
1276|         45|          26|Intermediate Level|         2 hours|2016-12-1
3 14:57:18|Business Finance|2016|
|   192870|Trading Penny Sto...|https://www.udemy...|   true|  150|
9221|        138|          25|        All Levels|         3 hours|2014-05-0
2 15:13:30|Business Finance|2014|
|   739964|Investing And Tra...|https://www.udemy...|   true|   65|
1540|        178|          26|    Beginner Level|          1 hour|2016-02-2
1 18:23:12|Business Finance|2016|
|   403100|Trading Stock Cha...|https://www.udemy...|   true|   95|
2917|        148|          23|        All Levels|       2.5 hours|2015-01-3
0 22:13:03|Business Finance|2015|
|   476268|Options Trading 3...|https://www.udemy...|   true|  195|
5172|         34|          38|      Expert Level|       2.5 hours|2015-05-2
8 00:14:03|Business Finance|2015|
|  1167710|The Only Investme...|https://www.udemy...|   true|  200|
827|         14|          15|        All Levels|          1 hour|2017-04-18
18:13:32|Business Finance|2017|
|   592338|Forex Trading Sec...|https://www.udemy...|   true|  200|
4284|         93|          76|        All Levels|         5 hours|2015-09-1
1 16:47:02|Business Finance|2015|
|   975046|Trading Options W...|https://www.udemy...|   true|  200|
1380|         42|          17|        All Levels|          1 hour|2016-10-1
8 22:52:31|Business Finance|2016|
|   742602|Financial Managem...|https://www.udemy...|   true|   30|
3607|         21|          19|        All Levels|       1.5 hours|2016-02-0
3 18:04:01|Business Finance|2016|
|   794151|Forex Trading Cou...|https://www.udemy...|   true|  195|
4061|         52|          16|        All Levels|         2 hours|2016-03-1
6 15:40:19|Business Finance|2016|
|  1196544|Python Algo Tradi...|https://www.udemy...|   true|  200|
294|         19|          42|        All Levels|         7 hours|2017-04-28
16:41:44|Business Finance|2017|
```

```
|   504036|Short Selling: Le...|https://www.udemy...|   true|   75|
2276|        106|          19|Intermediate Level|       1.5 hours|2015-06-2
2 21:18:35|Business Finance|2015|
|   719698|Basic Technical A...|https://www.udemy...|   true|   20|
4919|         79|          16|    Beginner Level|       1.5 hours|2016-01-0
8 17:21:26|Business Finance|2016|
|   564966|The Complete Char...|https://www.udemy...|   true|  200|
2666|        115|          52|        All Levels|         4 hours|2015-08-1
0 21:07:35|Business Finance|2015|
|   606928|7 Deadly Mistakes...|https://www.udemy...|   true|   50|
5354|         24|          23|        All Levels|       1.5 hours|2015-09-2
1 18:10:34|Business Finance|2015|
|    58977|Financial Stateme...|https://www.udemy...|   true|   95|
8095|        249|          12|    Beginner Level|         35 mins|2013-06-0
9 00:21:26|Business Finance|2013|
+---------+--------------------+--------------------+-------+-----+--------
-------+-----------+------------+------------------+----------------+------
-------------+----------------+----+
only showing top 20 rows
```

In [49]:
```
df_year = df_date.groupBy('year').mean().select(['year','avg(price)'])
```

In [50]:
```
df_year = df_year.withColumnRenamed('avg(price)', 'average_price')
```

In [51]:
```
df_year.show()
```

```
+----+------------------+
|year|     average_price|
+----+------------------+
|2015| 71.19622245540398|
|2013|59.435483870967744|
|null|              null|
|2014| 54.21945701357466|
|2012| 44.75609756097561|
|2016| 75.90990990990991|
|2011|              62.0|
|2017|  84.5631825273011|
+----+------------------+
```

In [52]:
```
df_year = df_year.select(['year', format_number('average_price', 0)])
```

In [53]:
```
df_year.show()
```

```
+----+------------------------------+
|year|format_number(average_price, 0)|
+----+------------------------------+
|2015|                            71|
|2013|                            59|
|null|                          null|
|2014|                            54|
|2012|                            45|
|2016|                            76|
|2011|                            62|
|2017|                            85|
+----+------------------------------+
```

In [54]:
```
df_year.withColumnRenamed('format_number(average_price, 0)', 'average_price')
```

```
+----+-------------+
|year|average_price|
+----+-------------+
|2015|           71|
|2013|           59|
|null|         null|
|2014|           54|
|2012|           45|
|2016|           76|
|2011|           62|
|2017|           85|
+----+-------------+
```

In [55]:
```
df_year.orderBy('year').show()
```

```
+----+------------------------------+
|year|format_number(average_price, 0)|
+----+------------------------------+
|null|                          null|
|2011|                            62|
|2012|                            45|
|2013|                            59|
|2014|                            54|
|2015|                            71|
|2016|                            76|
|2017|                            85|
+----+------------------------------+
```

Thank You!!