```python
#!/usr/bin/env python
# coding: utf-8


################################################################################
# Código base: DataCamp
################################################################################


################################################################################
import pyspark
from pyspark.sql import functions as f

tLinea = "*"*80

################################################################################
def hagaAlto(pMensaje):
    print(tLinea)
    print(pMensaje)
    print(tLinea)

```

```
22  ##########################################################################################
23  from pyspark.sql import SparkSession
24  spark = SparkSession.builder.master("local[*]").appName('PySpark_Tutorial').getOrCreate()
25
26  ##########################################################################################
27  # cargar archivo
28  b_data = spark.read.csv(
29      'data/stocks_price_final.csv',
30      sep = ',',
31      header = True,
32      )
33
34  b_data.printSchema()
35
36  hagaAlto("05-esquema 1")
37
```

```
root
 |-- _c0: string (nullable = true)
 |-- symbol: string (nullable = true)
 |-- date: string (nullable = true)
 |-- open: string (nullable = true)
 |-- high: string (nullable = true)
 |-- low: string (nullable = true)
 |-- close: string (nullable = true)
 |-- volume: string (nullable = true)
 |-- adjusted: string (nullable = true)
 |-- market.cap: string (nullable = true)
 |-- sector: string (nullable = true)
 |-- industry: string (nullable = true)
 |-- exchange: string (nullable = true)

**************************************************

05-esquema 1
**************************************************
```

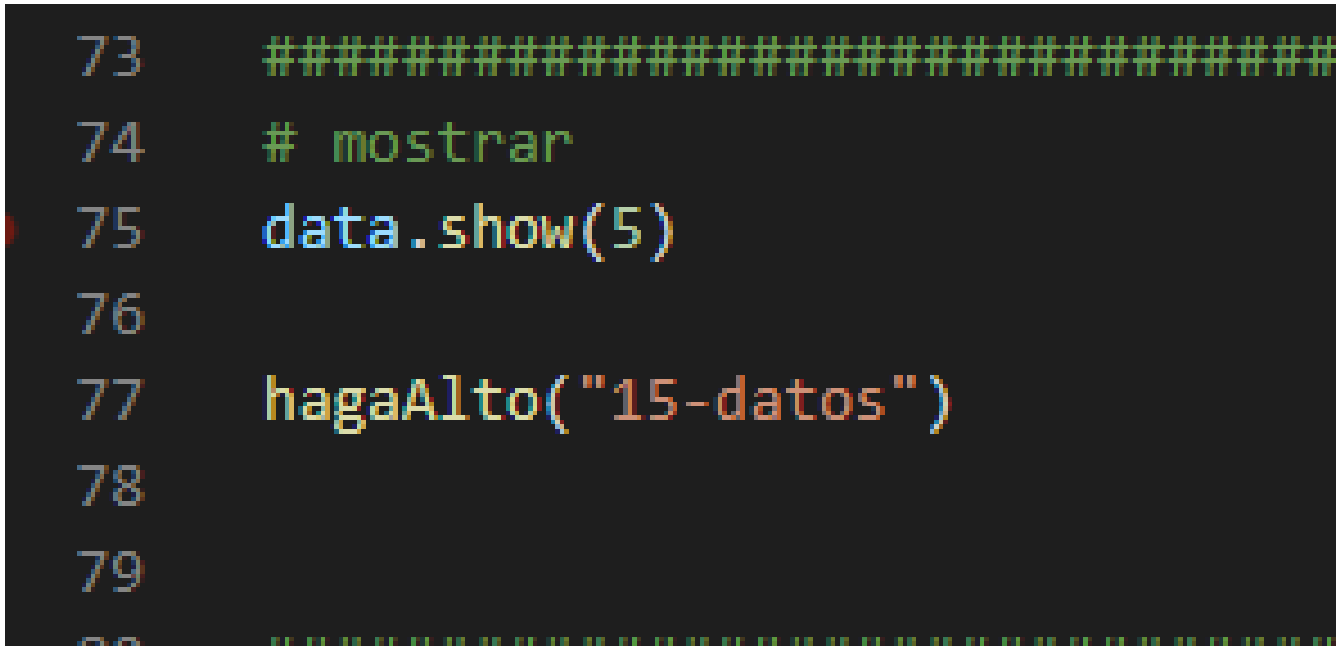
```
38  ##########################################################################
39  # cambiar estructura
40  from pyspark.sql.types import *
41
42  data_schema = [
43                  StructField('_c0', IntegerType(), True),
44                  StructField('symbol', StringType(), True),
45                  StructField('data', DateType(), True),
46                  StructField('open', DoubleType(), True),
47                  StructField('high', DoubleType(), True),
48                  StructField('low', DoubleType(), True),
49                  StructField('close', DoubleType(), True),
50                  StructField('volume', IntegerType(), True),
51                  StructField('adjusted', DoubleType(), True),
52                  StructField('market.cap', StringType(), True),
53                  StructField('sector', StringType(), True),
54                  StructField('industry', StringType(), True),
55                  StructField('exchange', StringType(), True),
56              ]
57
58  final_struc = StructType(fields=data_schema)
59
```

```
60  ##############################################
61  # se lee con la estructura
62  data = spark.read.csv(
63      'data/stocks_price_final.csv',
64      sep = ',',
65      header = True,
66      schema = final_struc
67      )
68
69  data.printSchema()
70
71  hagaAlto("10-esquema 2")
72
```

```
22/03/31 03:10:40 INFO InMemoryFileIndex: It took
root
 |-- _c0: integer (nullable = true)
 |-- symbol: string (nullable = true)
 |-- data: date (nullable = true)
 |-- open: double (nullable = true)
 |-- high: double (nullable = true)
 |-- low: double (nullable = true)
 |-- close: double (nullable = true)
 |-- volume: integer (nullable = true)
 |-- adjusted: double (nullable = true)
 |-- market.cap: string (nullable = true)
 |-- sector: string (nullable = true)
 |-- industry: string (nullable = true)
 |-- exchange: string (nullable = true)

**************************************************

10-esquema 2
**************************************************
```


73 ##
74 # mostrar
75 data.show(5)
76
77 hagaAlto("15-datos")
78
79

```
+---+------+----------+---------+---------+---------+---------+-------+---------+----------+------------+--------------------+--------+
|_c0|symbol|      data|     open|     high|      low|    close| volume| adjusted|market.cap|      sector|            industry|exchange|
+---+------+----------+---------+---------+---------+---------+-------+---------+----------+------------+--------------------+--------+
|  1|   TXG|2019-09-12|     54.0|     58.0|     51.0|    52.75|7326300|    52.75|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  2|   TXG|2019-09-13|    52.75|   54.355|49.150002|    52.27|1025200|    52.27|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  3|   TXG|2019-09-16|52.450001|     56.0|52.009998|55.200001| 269900|55.200001|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  4|   TXG|2019-09-17|56.209999|60.900002|   55.423|56.779999| 602800|56.779999|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  5|   TXG|2019-09-18|56.849998|    62.27|55.650002|     62.0|1589600|     62.0|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
+---+------+----------+---------+---------+---------+---------+-------+---------+----------+------------+--------------------+--------+
only showing top 5 rows

**********************************************************************************
15-datos
**********************************************************************************
```

```
80  ##########################################################################
81  # manejo de columnas
82  data = data.withColumnRenamed('market.cap', 'market_cap')
83
84  data = data.withColumn('date', data.data)
85
86  data.show(5)
87
88  hagaAlto("20-show 2")
89
90
```

_c0	symbol	data	open	high	low	close	volume	adjusted	market_cap	sector	industry	exchange	date
1	TXG	2019-09-12	54.0	58.0	51.0	52.75	7326300	52.75	$9.31B	Capital Goods	Biotechnology: La...	NASDAQ	2019-09-12
2	TXG	2019-09-13	52.75	54.355	49.150002	52.27	1025200	52.27	$9.31B	Capital Goods	Biotechnology: La...	NASDAQ	2019-09-13
3	TXG	2019-09-16	52.450001	56.0	52.009998	55.200001	269900	55.200001	$9.31B	Capital Goods	Biotechnology: La...	NASDAQ	2019-09-16
4	TXG	2019-09-17	56.209999	60.900002	55.423	56.779999	602800	56.779999	$9.31B	Capital Goods	Biotechnology: La...	NASDAQ	2019-09-17
5	TXG	2019-09-18	56.849998	62.27	55.650002	62.0	1589600	62.0	$9.31B	Capital Goods	Biotechnology: La...	NASDAQ	2019-09-18

only showing top 5 rows

****

20-show 2

****

```
91  ##########################################################################
92  # manejo de columnas
93  data = data.withColumnRenamed('date', 'data_changed')
94
95  data.show(5)
96
97  hagaAlto("25-show 3")
98
```

```
+---+------+----------+---------+---------+---------+---------+------+---------+----------+-------------+--------------------+--------+------------+
|_c0|symbol|      data|     open|     high|      low|    close|volume|adjusted|market_cap|       sector|            industry|exchange|data_changed|
+---+------+----------+---------+---------+---------+---------+------+---------+----------+-------------+--------------------+--------+------------+
|  1|   TXG|2019-09-12|     54.0|     58.0|     51.0|    52.75|7326300|    52.75|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|  2019-09-12|
|  2|   TXG|2019-09-13|    52.75|   54.355|49.150002|    52.27|1025200|    52.27|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|  2019-09-13|
|  3|   TXG|2019-09-16|52.450001|     56.0|52.009998|55.200001| 269900|55.200001|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|  2019-09-16|
|  4|   TXG|2019-09-17|56.209999|60.900002|   55.423|56.779999| 602800|56.779999|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|  2019-09-17|
|  5|   TXG|2019-09-18|56.849998|    62.27|55.650002|     62.0|1589600|     62.0|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|  2019-09-18|
+---+------+----------+---------+---------+---------+---------+------+---------+----------+-------------+--------------------+--------+------------+
only showing top 5 rows

********************************************************************************
25-show 3
********************************************************************************
```

```
 99    ##########################################
100    # borrar una columna
101    data = data.drop('data_changed')
102
103    data.show(5)
104
105    hagaAlto("30-luego de drop")
106
107
```

```
+---+------+----------+---------+---------+---------+---------+-------+---------+----------+-------------+--------------------+--------+
|_c0|symbol|      data|     open|     high|      low|    close| volume| adjusted|market_cap|       sector|            industry|exchange|
+---+------+----------+---------+---------+---------+---------+-------+---------+----------+-------------+--------------------+--------+
|  1|   TXG|2019-09-12|     54.0|     58.0|     51.0|    52.75|7326300|    52.75|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  2|   TXG|2019-09-13|    52.75|   54.355|49.150002|    52.27|1025200|    52.27|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  3|   TXG|2019-09-16|52.450001|     56.0|52.009998|55.200001| 269900|55.200001|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  4|   TXG|2019-09-17|56.209999|60.900002|   55.423|56.779999| 602800|56.779999|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  5|   TXG|2019-09-18|56.849998|    62.27|55.650002|     62.0|1589600|     62.0|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
+---+------+----------+---------+---------+---------+---------+-------+---------+----------+-------------+--------------------+--------+
only showing top 5 rows

**********************************************************************************
30-luego de drop
**********************************************************************************
```

```
108    ##########################################################################
109    # datos faltantes
110    data.na.drop()
111
112    data.na.fill(data.select(f.mean(data['open'])).collect()[0][0])
113
114
```

```
115  ##################################################
116  # ## 5. Selección de datos con PySpark SQL
117  # * Select
118  # * Filter
119  # * Between
120  # * When
121  # * Like
122  # * GroupBy
123  # * Aggregations
124
```

```
125    ##########################################################################
126    data.select(['open', 'high', 'low', 'close', 'volume', 'adjusted']).describe().show()
127
128    hagaAlto("35-luego de llenar missing values")
129
```

summary	open	high	low	close	volume	adjusted
count	1726301	1726301	1726301	1726301	1725207	1726301
mean	15070.071703341051	15555.06726813709	14557.808227578982	15032.714854330707	1397692.1627885813	14926.1096887955
stddev	1111821.8002863196	1148247.1953514954	1072968.1558434265	1109755.9294000647	5187522.908169119	1101877.6328940107
min	0.072	0.078	0.052	0.071	0	-1.230099
max	1.60168176E8	1.61601456E8	1.55151728E8	1.58376592E8	656504200	1.57249392E8

****

35-luego de llenar missing values

****

```
130  ##################################################################
131  from pyspark.sql.functions import col, lit
132
133  data.filter( (col('data') >= lit('2020-01-01'))
134              & (col('data') <= lit('2020-01-31')) ).show(5)
135
136  hagaAlto("40-filtro data")
137
```

```
+---+------+----------+---------+---------+---------+---------+------+---------+----------+-------------+--------------------+--------+
|_c0|symbol|      data|     open|     high|      low|    close|volume| adjusted|market_cap|       sector|            industry|exchange|
+---+------+----------+---------+---------+---------+---------+------+---------+----------+-------------+--------------------+--------+
| 78|   TXG|2020-01-02|76.910004|77.989998|71.480003|72.830002|220200|72.830002|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
| 79|   TXG|2020-01-03|71.519997|76.188004|70.580002|75.559998|288300|75.559998|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
| 80|   TXG|2020-01-06|75.269997|77.349998|73.559998|75.550003|220600|75.550003|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
| 81|   TXG|2020-01-07|     76.0|77.279999|    75.32|75.980003|182400|75.980003|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
| 82|   TXG|2020-01-08|76.089996|76.949997|72.739998|74.839996|172100|74.839996|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
+---+------+----------+---------+---------+---------+---------+------+---------+----------+-------------+--------------------+--------+
only showing top 5 rows

**********************************************************************************
40-filtro data
**********************************************************************************
```

```
138    ##################################################################
139    data.filter(data.adjusted.between(100.0, 500.0)).show(5)
140
141    hagaAlto("45-filtro between")
142
143
```

```
+----+------+----------+----------+----------+----------+----------+------+----------+----------+------------+--------------------+--------+
| _c0|symbol|      data|      open|      high|       low|     close|volume|  adjusted|market_cap|      sector|            industry|exchange|
+----+------+----------+----------+----------+----------+----------+------+----------+----------+------------+--------------------+--------+
|  93|   TXG|2020-01-24| 95.459999|     101.0| 94.157997|100.790001|328100|100.790001|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  94|   TXG|2020-01-27| 99.760002|104.892998| 97.019997|103.209999|334900|103.209999|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|  95|   TXG|2020-01-28|104.620003|108.269997|103.297997|106.620003|245400|106.620003|    $9.31B|Capital Goods|Biotechnology: La...|  NASDAQ|
|6893|  ABMD|2019-01-02|315.940002|320.709991|307.029999|309.959991|590000|309.959991|   $13.39B|  Health Care|Medical/Dental In...|  NASDAQ|
|6894|  ABMD|2019-01-03|    307.25| 311.73999|293.660004|302.290009|665300|302.290009|   $13.39B|  Health Care|Medical/Dental In...|  NASDAQ|
+----+------+----------+----------+----------+----------+----------+------+----------+----------+------------+--------------------+--------+
only showing top 5 rows

**********************************************************************************
45-filtro between
**********************************************************************************
```

```
144  ##########################################################################
145  data.select('open', 'close',
146                  f.when(data.adjusted >= 200.0, 1).otherwise(0)).show(5)
147
148  hagaAlto("50-filtro ajusted")
149
```

```
+---------+---------+-----------------------------------------------+
|     open|    close|CASE WHEN (adjusted >= 200.0) THEN 1 ELSE 0 END|
+---------+---------+-----------------------------------------------+
|     54.0|    52.75|                                              0|
|    52.75|    52.27|                                              0|
|52.450001|55.200001|                                              0|
|56.209999|56.779999|                                              0|
|56.849998|     62.0|                                              0|
+---------+---------+-----------------------------------------------+
only showing top 5 rows

**************************************************************************************
50-filtro ajusted
**************************************************************************************
```

```
150  ##########################################################################################
151  data.select('sector',
152              data.sector.rlike('^[B,C]').alias('Sector Starting with B or C')
153              ).distinct().show()
154
155  hagaAlto("55-iniciando con B o C")
156
```

```
+--------------------+---------------------------+
|              sector|Sector Starting with B or C|
+--------------------+---------------------------+
|         Health Care|                      false|
|       Capital Goods|                       true|
|Consumer Non-Dura...|                       true|
|    Public Utilities|                      false|
|   Consumer Durables|                       true|
|             Finance|                      false|
|      Transportation|                      false|
|       Miscellaneous|                      false|
|   Consumer Services|                       true|
|              Energy|                      false|
|    Basic Industries|                       true|
|          Technology|                      false|
+--------------------+---------------------------+


**************************************************
55-iniciando con B o C
**************************************************
```

```
157  ##############################################################################
158  data.select(['industry', 'open', 'close', 'adjusted']).groupBy('industry').mean().show()
159  hagaAlto("60-media por industria")
160
```

industry	avg(open)	avg(close)	avg(adjusted)
Finance/Investors...	5.134401785714286	5.136630739795919	4.991354066964286
Miscellaneous	16.38588266938776	16.35987909030613	16.148959322959183
Investment Banker...	58.95058094575029	58.983085960826294	58.157837258903065
Food Distributors	43.274508569354644	43.27317810574859	42.910476083578644
Miscellaneous man...	15.660586409948984	15.65093486096939	15.369818847193866
Ophthalmic Goods	108.50137892138572	108.54045987608258	108.52516121052633
Broadcasting	24.916787464825223	24.91738845539514	24.699102029625255
Agricultural Chem...	22.046413928996614	22.042051076318053	21.635093418154767
Biotechnology: Bi...	24.808083192324542	24.803587149935442	24.74507997827319
Other Specialty S...	84.80718810562882	84.80276550929834	84.55525036482379
Biotechnology: El...	33.36891734535045	33.33611913546892	33.21022605613573
Other Consumer Se...	43.67010744224583	43.658688711464606	43.4349898087902
Electric Utilitie...	41.35569183903091	41.37105559357328	40.39245735242015
Specialty Foods	65.22351357692312	65.22317585370249	64.18661875197694
Plastic Products	31.69500596129026	31.70773089354834	31.40776585092165
Precision Instrum...	24.476071367346933	24.506250015306108	24.506250015306108
Water Supply	40.5804830820354	40.58487374462944	40.17332791487649
Banks	21.441229607680004	21.440168331039978	21.029210073439987
Farming/Seeds/Mil...	27.74014344411733	27.74297949099047	27.022539238958878
Medical/Nursing S...	71.0372895288078	71.09947781274889	70.77964412074775

only showing top 20 rows

****

60-media por industria

****

```
161  ##########################################################################
162  from pyspark.sql.functions import col, min, max, avg, lit
163
164  data.groupBy("sector").agg(min("data").alias("From"),
165              max("data").alias("To"),
166              min("open").alias("Minimum Opening"),
167              max("open").alias("Maximum Opening"),
168              avg("open").alias("Average Opening"),
169              min("close").alias("Minimum Closing"),
170              max("close").alias("Maximum Closing"),
171              avg("close").alias("Average Closing"),
172              min("adjusted").alias("Minimum Adjusted Closing"),
173              max("adjusted").alias("Maximum Adjusted Closing"),
174              avg("adjusted").alias("Average Adjusted Closing"),
175          ).show(truncate=False)
176
177  hagaAlto("65-estadísticas")
178
```

sector	From	To	Minimum Opening	Maximum Opening	Average Opening	Minimum Closing	Maximum Closing	Maximum Adjusted Closing	Average Adjusted Closing
Miscellaneous	2019-01-02	2020-07-22	0.147	1059.98999	52.03839496900624	0.1361	1035.829956	1035.829956	51.809730336322794
Health Care	2019-01-02	2020-07-22	0.072	186000.0	119.96763306523218	0.071	187000.0	187000.0	118.97394778016198
Public Utilities	2019-01-02	2020-07-22	0.331	280.0	35.580777352394705	0.325	282.220001	280.67395	34.7301556850045
Energy	2019-01-02	2020-07-22	0.1	905.0	24.456589891261007	0.09	901.039978	879.057007	23.684714263001084
Consumer Non-Durables	2019-01-02	2020-07-22	0.12	655.0	43.32860274612677	0.12	664.130005	664.130005	42.81762456569033
Finance	2019-01-02	2020-07-22	0.25	1336.930054	37.77466706818995	0.27	1341.079956	1341.079956	37.10028522718097
Basic Industries	2019-01-02	2020-07-22	0.23	1.60168176E8	266410.35470107093	0.23	1.58376592E8	1.57249392E8	263865.51070311887
Capital Goods	2019-01-02	2020-07-22	0.13	4025.0	60.4885436328285	0.12	4037.77002	4037.77002	59.97512253879351
Technology	2019-01-02	2020-07-22	0.14	2704.0	49.516045118395034	0.13	2736.0	2736.0	49.25234033754425
Consumer Services	2019-01-02	2020-07-22	0.1	15437.5	55.078867342590755	0.134	19843.75	19843.75	54.36187827846176
Consumer Durables	2019-01-02	2020-07-22	0.32	111718.75	391.03153998497794	0.31	118750.0	118750.0	389.6417639493677
Transportation	2019-01-02	2020-07-22	0.08	274.410004	37.30503242702824	0.08	274.040009	274.040009	36.80231922791236

****
65-estadísticas
****

```
179  ###############################################################################
180  data.filter( (col('data') >= lit('2019-01-02')) & (col('data') <= lit('2020-01-31')) ).groupBy("sector").agg(min("data").alias("From"),
181          max("data").alias("To"),
182          min("open").alias("Minimum Opening"),
183          max("open").alias("Maximum Opening"),
184          avg("open").alias("Average Opening"),
185          min("close").alias("Minimum Closing"),
186          max("close").alias("Maximum Closing"),
187          avg("close").alias("Average Closing"),
188
189          min("adjusted").alias("Minimum Adjusted Closing"),
190          max("adjusted").alias("Maximum Adjusted Closing"),
191          avg("adjusted").alias("Average Adjusted Closing"),
192
193      ).show(truncate=False)
194  hagaAlto("70-filtros")
195
```

sector	From	To	Minimum Opening	Maximum Opening	Average Opening	Minimum Closing	Maximum Closing	Maximum Adjusted Closing	Average Adjusted Closing
Miscellaneous	2019-01-02	2020-01-31	0.16	690.0	51.51619596530094	0.16	691.099976	691.099976	51.21050733851903
Health Care	2019-01-02	2020-01-31	0.072	186000.0	146.77561975676196	0.071	187000.0	187000.0	145.50092663131548
Public Utilities	2019-01-02	2020-01-31	0.62	269.459991	36.67104182710834	0.66	269.929993	267.056549	35.55794068386673
Energy	2019-01-02	2020-01-31	0.13	905.0	27.695591382025718	0.13	901.039978	879.057007	26.675108967542013
Consumer Non-Durables	2019-01-02	2020-01-31	0.22	441.190002	44.68703179837432	0.22	441.26001	441.26001	44.00786604732114
Finance	2019-01-02	2020-01-31	0.284	1199.030029	40.304979730179085	0.27	1196.22998	1196.22998	39.41408818854872
Basic Industries	2019-01-02	2020-01-31	0.277	1.60168176E8	334814.23540859646	0.272	1.58376592E8	1.57249392E8	331916.3926153813
Capital Goods	2019-01-02	2020-01-31	0.21	4025.0	61.479283472745024	0.22	4030.0	4030.0	60.79992564173739
Technology	2019-01-02	2020-01-31	0.14	2704.0	49.63989857764513	0.13	2736.0	2736.0	49.271465831557585
Consumer Services	2019-01-02	2020-01-31	0.21	15437.5	58.50570512458658	0.201	19843.75	19843.75	57.56926626388146
Consumer Durables	2019-01-02	2020-01-31	0.32	106250.0	439.96307151506704	0.31	103906.25	103906.25	436.1655366019183
Transportation	2019-01-02	2020-01-31	0.61	268.190002	39.514098111525676	0.63	269.630005	267.881714	38.86532091563174

****
70-filtros
****

```
196  ##############################################################################
197  # Escribir a archivos
198  # CSV
199  data.write.csv('dataset.csv')
200
201  # JSON
202  data.write.save('dataset.json', format='json')
203
204  ## Writing selected data to different file formats
205
206  # CSV
207  data.select(['data', 'open', 'close', 'adjusted']).write.csv('dataset2.csv')
208
209  # JSON
210  data.select(['data', 'open', 'close', 'adjusted']).write.save('dataset2.json', format='json')
211
212  hagaAlto("75-FIN DEL PROGRAMA")
213  ##############################################################################
214  # fin
215  ##############################################################################
216
```


75-FIN DEL PROGRAMA