Punto3

Head default

```
%pyspark
import pyspark
from pyspark.sql.session import SparkSession
from pyspark.sql.functions import *
```

FINISHED

Took 36 sec. Last updated by anonymous at March 31 2019, 3:53:48 PM.

```
%pyspark
# creamos la sesión de spark
spark = SparkSession\
.builder\
.appName("spark_Parte3")\
.getOrCreate()
```

FINISHED

Took 1 sec. Last updated by anonymous at March 31 2019, 3:53:54 PM.

```
%pyspark
profecoDf = spark.read.format("csv")\
        .option("delimiter", "|")\
        .option("header","true")\
        .option("inferSchema", "true")\
        .load("s3://urielmminonawsbucket/profeco_final_bash.csv")
```

SPARK JOBS FINISHED

Took 10 min 10 sec. Last updated by anonymous at March 31 2019, 4:04:28 PM.

```
%pyspark
profecoDf.columns
```

FINISHED

```
['producto', 'presentacion', 'marca', 'categoria', 'catalogo', 'precio', 'fecharegistro', 'cadenacomercial', 'giro', 'nombrecomercial', 'direccion', 'estado', 'municipio', 'latitud', 'longitud']
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:13:02 PM.

```
%pyspark
profecoDf.count()
```

SPARK JOB FINISHED

```
62530715
```

Took 1 min 30 sec. Last updated by anonymous at March 31 2019, 4:14:41 PM.

```
%pyspark
profecoDf.show(5)
```

SPARK JOB FINISHED

```
|            producto|        presentacion|   marca|       categoria|        catalogo|precio|      fecharegistro|   cadenacomercial|      giro|     nombrecomercial|           direccion|          estado|          muni
cipio|latitud| longitud|
+--------------------+--------------------+--------+----------------+----------------+------+-------------------+------------------+----------+--------------------+--------------------+----------------+--------------
-----+-------+---------+
|cuaderno forma it...|96 hojas pasta du...|estrella|material escolar|utiles escolares|   259|2011-05-18 00:00:00|abastecedora lumen|papelerias|abastecedora lume...|cannes no 6 esq c...|distrito federal|tlalpan
...|1929699|-99125417|
|            crayones|caja 12 ceras jum...| crayola|material escolar|utiles escolares|   275|2011-05-18 00:00:00|abastecedora lumen|papelerias|abastecedora lume...|cannes no 6 esq c...|distrito federal|tlalpan
...|1929699|-99125417|
|            crayones|caja 12 ceras tam...| crayola|material escolar|utiles escolares|   139|2011-05-18 00:00:00|abastecedora lumen|papelerias|abastecedora lume...|cannes no 6 esq c...|distrito federal|tlalpan
...|1929699|-99125417|
|   colores de madera|caja 12 piezas la...|pincelin|material escolar|utiles escolares|   469|2011-05-18 00:00:00|abastecedora lumen|papelerias|abastecedora lume...|cannes no 6 esq c...|distrito federal|tlalpan
...|1929699|-99125417|
|         color largo|caja 36 piezas co...| crayola|material escolar|utiles escolares|   115|2011-05-18 00:00:00|abastecedora lumen|papelerias|abastecedora lume...|cannes no 6 esq c...|distrito federal|tlalpan
...|1929699|-99125417|
+--------------------+--------------------+--------+----------------+----------------+------+-------------------+------------------+----------+--------------------+--------------------+----------------+--------------
-----+-------+---------+
only showing top 5 rows
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:14:45 PM.

```
%pyspark
profecoDf.dtypes
```

FINISHED

```
[('producto', 'string'), ('presentacion', 'string'), ('marca', 'string'), ('categoria', 'string'), ('catalogo', 'string'), ('precio', 'int'), ('fecharegistro', 'timestamp'), ('cadenacomercial', 'string'), ('giro', 'st
ring'), ('nombrecomercial', 'string'), ('direccion', 'string'), ('estado', 'string'), ('municipio', 'string'), ('latitud', 'string'), ('longitud', 'string')]
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:14:50 PM.

3.2 Guardar como .parquet

```
%pyspark
profecoDf.write.parquet('s3://urielmminonawsbucket/profecoFinal.parquet')
```

SPARK JOB FINISHED

Took 9 min 45 sec. Last updated by anonymous at March 31 2019, 4:24:39 PM.

3.3 Cargar parquet en Spark

```
%pyspark
profecoDfParquet = spark.read.format('parquet')\
        .load("s3://urielmminonawsbucket/profecoFinal.parquet")
```

SPARK JOB FINISHED

Took 0 sec. Last updated by anonymous at March 31 2019, 4:37:08 PM.

```
%pyspark
profecoDfParquet.count()
```

SPARK JOB FINISHED

```
62530715
```

Took 4 sec. Last updated by anonymous at March 31 2019, 4:37:14 PM.

```
%pyspark
profecoDfParquet.show(1)
```

SPARK JOB FINISHED

```
+--------------------+--------------------+-----+--------------------+----------------+------+-------------------+--------------+--------------------+--------------------+--------------------+--------+---------+---
-----+----------+
|            producto|        presentacion|marca|           categoria|        catalogo|precio|      fecharegistro|cadenacomercial|                giro|     nombrecomercial|           direccion|  estado|municipio| la
titud|  longitud|
+--------------------+--------------------+-----+--------------------+----------------+------+-------------------+--------------+--------------------+--------------------+--------------------+--------+---------+---
-----+----------+
|sistemas de teatr...|bh 4430 o bh 4430...|   lg|aparatos electron...|electrodomesticos|379900|2016-01-15 15:25:53|       elektra|tienda de electro...|elektra sucursal ...|boulevard lopez m...|zacatecas|zacatecas|227
67258|-102574067|
+--------------------+--------------------+-----+--------------------+----------------+------+-------------------+--------------+--------------------+--------------------+--------------------+--------+---------+---
-----+----------+
only showing top 1 row
```

Took 1 sec. Last updated by anonymous at March 31 2019, 4:37:18 PM.

```
%pyspark
#Poner disponible para ejecutar sentencias SQL en Spark
profecoDfParquet.createOrReplaceTempView("profecoDf")
```

FINISHED

Took 0 sec. Last updated by anonymous at March 31 2019, 4:51:37 PM.

```
%pyspark
print(profecoDf.columns)
```

FINISHED

```
['producto', 'presentacion', 'marca', 'categoria', 'catalogo', 'precio', 'fecharegistro', 'cadenacomercial', 'giro', 'nombrecomercial', 'direccion', 'estado', 'municipio', 'latitud', 'longitud']
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:51:40 PM.

```
%pyspark
# Filtrar unicamente medicamentos
spark.sql("SELECT * FROM profecoDf WHERE categoria='medicamentos'").createOrReplaceTempView("profecoDfMed")
```

FINISHED

Took 0 sec. Last updated by anonymous at March 31 2019, 4:51:44 PM.

```
%pyspark
# Conteo de medicamentos
spark.sql("SELECT count(1) FROM profecoDfMed").show()
```

SPARK JOB FINISHED

```
+--------+
|count(1)|
+--------+
|11485813|
+--------+
```

Took 29 sec. Last updated by anonymous at March 31 2019, 4:52:18 PM.

```
%pyspark
# Existen productos sin marca identificados como 's/m' y 'sin marca' que representan una gran parte de los datos
spark.sql("SELECT count(1) FROM profecoDfMed WHERE marca IN ('s/m','sin marca')").show()
```

SPARK JOB FINISHED

```
+--------+
|count(1)|
+--------+
|10766552|
+--------+
```

Took 9 sec. Last updated by anonymous at March 31 2019, 4:52:41 PM.

```
%pyspark
spark.sql("SELECT DISTINCT YEAR(fecharegistro) FROM profecoDfMed WHERE marca IN ('s/m','sin marca')").show()
```

SPARK JOBS FINISHED

```
+---------------------------------+
|year(CAST(fecharegistro AS DATE))|
+---------------------------------+
|                             2015|
|                             2013|
|                             2014|
|                             2012|
|                             2016|
|                             2011|
+---------------------------------+
```

Took 13 sec. Last updated by anonymous at March 31 2019, 4:52:59 PM.

```
%pyspark
#Los productos sin marca corresponden a marcas genéricas
t1=spark.sql("SELECT DISTINCT nombrecomercial FROM profecoDfMed WHERE marca IN ('s/m','sin marca')")
z.show(t1)
```

SPARK JOBS FINISHED

settings

nombrecomercial
super san francisco de asis sucursal crucero
heb sucursal lincoln
bodega aurrera sucursal estadio
s-mart sucursal anzalduas
imss sucursal tequesquinahuac
farmacia vista hermosa
aurrera bodega sucursal gobernadora
farmacias del ahorro sucursal chiapas

Took 10 sec. Last updated by anonymous at March 31 2019, 4:53:13 PM.

```
%pyspark
# Se eliminan las marcas s/m y sin marca ya que son la mayoría y afecta los resultados del análisis
# spark.sql("SELECT * FROM profecoDfMed WHERE marca NOT IN ('s/m','sin marca')").createOrReplaceTempView("profecoDfMed")
```

FINISHED

Took 0 sec. Last updated by anonymous at March 31 2019, 4:27:43 PM. (outdated)

```
%pyspark
spark.sql("SELECT count(1) FROM profecoDfMed").show()
```

SPARK JOB FINISHED

```
+--------+
|count(1)|
+--------+
|11485813|
+--------+
```

Took 4 sec. Last updated by anonymous at March 31 2019, 4:53:24 PM.

```
%pyspark
# Eliminar Vista para liberar RAM
spark.catalog.dropTempView("profecoDf")
```

FINISHED

Took 0 sec. Last updated by anonymous at March 31 2019, 4:53:27 PM.

3.3 Pregunta 1 ¿Cuańtas marcas diferentes tiene tu categoría?

SPARK JOB FINISHED

```
%pyspark
p1 =spark.sql("SELECT COUNT(*) AS total FROM (SELECT DISTINCT(marca) FROM profecoDfMed)")
z.show(p1)
```

settings

total
77

3.3 Pregunta 2 ¿Cuál es la marca con mayor precio? ¿En qué estado?

SPARK JOB FINISHED

```
%pyspark
p2 = spark.sql("SELECT marca, precio, estado FROM profecoDfMed ORDER BY precio DESC LIMIT 1")
z.show(p2)
```

settings

marca	precio	estado
s/m	701978	mexico

Took 8 sec. Last updated by anonymous at March 31 2019, 4:53:51 PM. (outdated)

3.3 Pregunta 3 ¿Cuál es la marca con menor precio en CDMX? (en aquel entonces Distrito Federal)

SPARK JOB FINISHED

```
%pyspark
p3=spark.sql("SELECT marca, precio, estado FROM profecoDfMed WHERE estado='distrito federal' ORDER BY precio ASC LIMIT 1")
z.show(p3)
```

settings

marca	precio	estado
s/m	1	distrito federal

Took 7 sec. Last updated by anonymous at March 31 2019, 4:54:03 PM. (outdated)

3.3 Pregunta 4 ¿Cuál es la marca con mayores observaciones?

SPARK JOB FINISHED

```
%pyspark
p4 =spark.sql("SELECT marca, COUNT(marca) AS observaciones FROM profecoDfMed GROUP BY marca ORDER BY COUNT(marca) DESC LIMIT 1")
z.show(p4)
```

settings

marca	observaciones
s/m	9472779

3.3 Pregunta 5 ¿Ha dejado de existir alguna marca durante los años que tienes? ¿Cuál? ¿Cuándo desapareció?

SPARK JOB FINISHED

```
%pyspark
# Obtener el año máximo
maxYears =spark.sql("SELECT MAX(year(fecharegistro)) AS year FROM profecoDfMed").rdd.flatMap(lambda x: x).collect()
maxYears=str(maxYears[0])
maxYears
```

```
'2016'
```

Took 6 sec. Last updated by anonymous at March 31 2019, 4:54:52 PM.

FINISHED

```
%pyspark
spark.sql("SELECT DISTINCT marca, MAX(year(fecharegistro)) AS maxYear  \
          FROM profecoDfMed \
          GROUP BY marca").createOrReplaceTempView("profecoDistMarcYear")
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:54:55 PM.

FINISHED

```
%pyspark
# Marca y año que desapareció
# Se considera que el año en que desapareció es el año siguiente al último registrado
p5=spark.sql("SELECT marca, maxYear+1 AS desYear FROM profecoDistMarcYear WHERE maxYear<{0}".format(maxYears))
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:55:02 PM.

SPARK JOBS FINISHED

```
%pyspark
z.show(p5)
```

settings

marca	desYear
scnering mexicana, sa de cv	2013
farmacom - gi quimica y farmacia	2012
productos roche, sa de cv	2013
primer nivel gi arlex	2013
primer nivel gi waldel	2013
primer nivel arlex	2013
best	2013
abbott laboratories de mexico, sa de cv	2013

SPARK JOB FINISHED

```
%pyspark
p5.count()
```

```
76
```

Took 9 sec. Last updated by anonymous at March 31 2019, 4:55:42 PM.

FINISHED

```
%pyspark
# Crear Vista profecoDistMarcYear
profecoDfParquet.createOrReplaceTempView("profecoDistMarcYear")
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:55:53 PM.

3.3 Pregunta 6 Genera una gráfica de serie de tiempo por estado para la marca con mayor precio -en todos los años-, donde el eje equis es el año y el eje ye es el precio máximo.

FINISHED

```
%pyspark
datos_grafica = spark.sql("SELECT MAX(precio), estado FROM profecoDfMed GROUP BY estado")
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:56:01 PM.

FINISHED

```
%pyspark
spark.sql("SELECT DISTINCT A.marca,A.estado,A.precio \
        FROM profecoDfMed A\
        INNER JOIN \
            (SELECT MAX(precio) AS precio, estado FROM profecoDfMed GROUP BY estado) B \
        ON A.estado = B.estado \
        AND A.precio = B.precio").createOrReplaceTempView("profecoDfMaxPrecio")
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:56:48 PM.

FINISHED

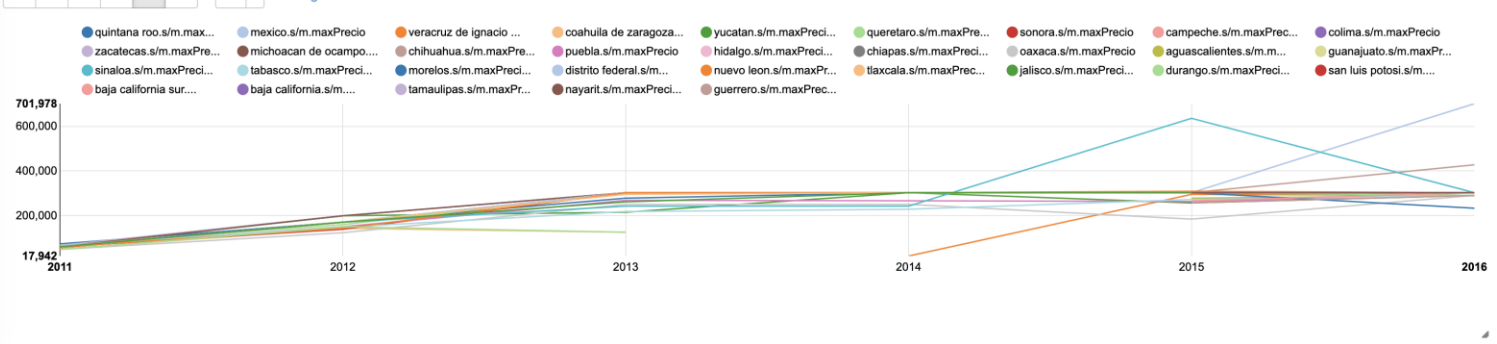
```
%pyspark
df = spark.sql("SELECT marca,estado,MAX(precio) as maxPrecio,YEAR(fecharegistro) as year \
        FROM profecoDfMed WHERE marca||estado IN (SELECT marca||estado FROM profecoDfMaxPrecio) \
        GROUP BY marca,estado,YEAR(fecharegistro)"
```

Took 0 sec. Last updated by anonymous at March 31 2019, 4:56:53 PM.

SPARK JOBS FINISHED

```
%pyspark
z.show(df)
```

settings



Took 1 min 0 sec. Last updated by anonymous at March 31 2019, 4:57:56 PM. (outdated)

READY

```
%pyspark
# Eliminar Vista para liberar RAM
spark.catalog.dropTempView("profecoDfMaxPrecio")
spark.catalog.dropTempView("profecoDfMed")
```