Jupyter Notebook
Ejercicio B Last Checkpoint: hace una hora (unsaved changes) Current Kernel Logo Logout
Menu

Python 3

Trusted

- File
  - New NotebookToggle Dropdown
    - Python 3
  - Open...
  -
  - Make a Copy...
  - Save as...
  - Rename...
  - Save and Checkpoint
  -
  - Revert to CheckpointToggle Dropdown
    - domingo, 16 de mayo de 2021 20:56
  -
  - Print Preview
  - Download asToggle Dropdown
    - asciidoc (.asciidoc)
    - html (.html)
    - latex (.tex)
    - markdown (.md)
    - notebook (.ipynb)
    - pdf (.pdf)
    - rst (.rst)
    - Python (.py)
    - slides (.slides.html)
  - Deploy as
  -
  - Trusted Notebook
  -
  - Close and Halt
- Edit
  - Cut Cells
  - Copy Cells
  - Paste Cells Above
  - Paste Cells Below
  - Paste Cells & Replace
  - Delete Cells
  - Undo Delete Cells
  -
  - Split Cell
  - Merge Cell Above
  - Merge Cell Below
  -
  - Move Cell Up
  - Move Cell Down
  -
  - Edit Notebook Metadata
  -
  - Find and Replace
  -
  - Cut Cell Attachments
  - Copy Cell Attachments
  - Paste Cell Attachments
  -
  - Insert Image
- View
  - Toggle Header
  - Toggle Toolbar
  - Toggle Line Numbers
  - Cell Toolbar
    - None
    - Edit Metadata
    - Raw Cell Format
    - Slideshow
    - Attachments
    - Tags
- Insert
  - Insert Cell Above
  - Insert Cell Below
- Cell
  - Run Cells
  - Run Cells and Select Below
  - Run Cells and Insert Below
  - Run All
  - Run All Above
  - Run All Below
  -
  - Cell Type
    - Code
    - Markdown
    - Raw NBConvert
  -
  - Current Outputs
    - Toggle
    - Toggle Scrolling
    - Clear
  - All Output
    - Toggle
    - Toggle Scrolling
    - Clear
- Kernel
  - Interrupt
  - Restart
  - Restart & Clear Output
  - Restart & Run All
  - Reconnect
  - Shutdown
  -
  - Change kernel
    - Python 3
- Widgets
  - Save Notebook Widget State
  - Clear Notebook Widget State
  - Download Widget State
  - Embed Widgets
- Help
  - User Interface Tour
  - Keyboard Shortcuts
  - Edit Keyboard Shortcuts
  -
  - Notebook Help
  - Markdown
  -
  - Python Reference
  - IPython Reference
  - NumPy Reference
  - SciPy Reference
  - Matplotlib Reference
  - SymPy Reference
  - pandas Reference
  -
  - About

Run
Code

In [26]:

x

```
import pyspark

from pyspark.sql import SparkSession

from pyspark.sql.functions import rank, col

from pyspark.sql.window import Window

spark = SparkSession \

    .builder \

    .appName("Python Spark SQL Opi") \

    .config("spark.some.config.option", "some-value") \

    .getOrCreate()
```

In [2]:

```
df = spark.read.csv("all_data.csv",header=True)
```

In [3]:

```
df.show()
```

```
+--------------------+--------------------+--------------------+--------------------+-----------------+------+--------------------+------------------+--------------------+--------------------+--------------------+----------------+--------------------+--------+----------+
|            producto|        presentacion|                marca|           categoria|         catalogo|precio|       fechaRegistro|   cadenaComercial|                giro|     nombreComercial|           direccion|          estado|           municipio| latitud|  longitud|
+--------------------+--------------------+--------------------+--------------------+-----------------+------+--------------------+------------------+--------------------+--------------------+--------------------+----------------+--------------------+--------+----------+
|CUADERNO FORMA IT...|96 HOJAS PASTA DU...|            ESTRELLA|    MATERIAL ESCOLAR| UTILES ESCOLARES|  25.9|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|            CRAYONES|CAJA 12 CERAS. JU...|             CRAYOLA|    MATERIAL ESCOLAR| UTILES ESCOLARES|  27.5|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|            CRAYONES|CAJA 12 CERAS. TA...|             CRAYOLA|    MATERIAL ESCOLAR| UTILES ESCOLARES|  13.9|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|   COLORES DE MADERA|CAJA 12 PIEZAS LA...|            PINCELIN|    MATERIAL ESCOLAR| UTILES ESCOLARES|  46.9|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|         COLOR LARGO|CAJA 36 PIEZAS. C...|             CRAYOLA|    MATERIAL ESCOLAR| UTILES ESCOLARES|   115|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|          BOLIGRAFO|BLISTER 3 PIEZAS....|    BIC. CRISTAL GEL|    MATERIAL ESCOLAR| UTILES ESCOLARES|  32.5|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|     CINTA ADHESIVA|BOLSA 1 PIEZA. 12...|     SCOTCH 3M. 600|    MATERIAL ESCOLAR| UTILES ESCOLARES|     9|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|   COLORES DE MADERA|CAJA 24 PIEZAS. L...|      KORES. KOLORES|    MATERIAL ESCOLAR| UTILES ESCOLARES|  95.9|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|            CRAYONES|CAJA 24 CERAS. TA...|  PAPER MATE. CARMEN|    MATERIAL ESCOLAR| UTILES ESCOLARES|  23.2|2011-05-18 00:00:...|ABASTECEDORA LUMEN|          PAPELERIAS|ABASTECEDORA LUME...|CANNES No. 6 ESQ....|DISTRITO FEDERAL|TLALPAN          ...|19.29699|-99.125417|
|  PAN BLANCO BOLILLO|               PIEZA|                 S/M|                 PAN|          BASICOS|   1.2|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|    HARINA HOT CAKES|       CAJA 800 GR.|PRONTO TRADICIONALES|GALLETAS PASTAS Y...|          BASICOS| 21.63|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|     PASTA PARA SOPA|PAQUETE 200 GR. S...|               VESTA|GALLETAS PASTAS Y...|          BASICOS|  3.45|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|     GALLETAS DULCES|CAJA 652 GR. RICA...|              GAMESA|GALLETAS PASTAS Y...|          BASICOS| 21.77|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|     GALLETAS DULCES|PAQUETE 280 GR. P...|            MARINELA|GALLETAS PASTAS Y...|          BASICOS|    13|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|     GALLETAS DULCES|CAJA 752 GR. SAND...|             NABISCO|GALLETAS PASTAS Y...|          BASICOS| 41.97|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|             SHAMPOO|BOTELLA 400 ML. P...|   HEAD & SHOULDERS|ARTS. PARA EL CUI...|          BASICOS|  49.9|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
| HORNO DE MICROONDAS|MS-1446 SQP O  MS...|                  LG| APARATOS ELECTRICOS|ELECTRODOMESTICOS|  1598|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|ACONDICIONADOR Y ...|BOTELLA 300 ML. F...|             GARNIER|ARTS. PARA EL CUI...|          BASICOS|  39.9|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|TINTE PARA EL CAB...|                CAJA|     REVITALIQUE. 4|ARTS. PARA EL CUI...|          BASICOS|    74|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
|TINTE PARA EL CAB...|                CAJA|  WELLA KOLESTON. 40|ARTS. PARA EL CUI...|          BASICOS|  58.9|2011-01-10 00:00:...|COMERCIAL MEXICANA|TIENDA DE AUTOSER...|COMERCIAL MEXICAN...|AV. LAGO DE GUADA...|          MÉXICO|ATIZAPAN         ...|      NA|        NA|
+--------------------+--------------------+--------------------+--------------------+-----------------+------+--------------------+------------------+--------------------+--------------------+--------------------+----------------+--------------------+--------+----------+
only showing top 20 rows
```

In [4]:

x

```
df.printSchema()
```

```
root
 |-- producto: string (nullable = true)
 |-- presentacion: string (nullable = true)
 |-- marca: string (nullable = true)
 |-- categoria: string (nullable = true)
 |-- catalogo: string (nullable = true)
 |-- precio: string (nullable = true)
 |-- fechaRegistro: string (nullable = true)
 |-- cadenaComercial: string (nullable = true)
 |-- giro: string (nullable = true)
 |-- nombreComercial: string (nullable = true)
 |-- direccion: string (nullable = true)
 |-- estado: string (nullable = true)
 |-- municipio: string (nullable = true)
 |-- latitud: string (nullable = true)
 |-- longitud: string (nullable = true)
```

In [5]:

```
# Pregunta 1_a: encontrando numero de registros en la base de datos

df.count()
```

Out[5]:

```
62530715
```

In [7]:

x

```
df.select("categoria").distinct().show()
```

```
+--------------------+
|           categoria|
+--------------------+
|    MATERIAL ESCOLAR|
|ARTS. PARA EL CUI...|
|PESCADOS Y MARISC...|
|UTENSILIOS DOMEST...|
|           categoria|
|DETERGENTES Y PRO...|
|CARNE Y VISCERAS ...|
|PRODUCTOS DE TEMP...|
|GALLETAS PASTAS Y...|
|  HORTALIZAS FRESCAS|
|                null|
|  DERIVADOS DE LECHE|
|TORTILLAS Y DERIV...|
|GRASAS ANIMALES C...|
|APARATOS ELECTRON...|
|     LEGUMBRES SECAS|
|                CAFE|
|   MUEBLES DE COCINA|
|CARNES FRIAS SECA...|
|CHOCOLATES Y GOLO...|
+--------------------+
only showing top 20 rows
```

In [8]:

x

```
#Pregunta 1_b: encontrando el numero de categorias

df.select("categoria").distinct().count()
```

Out[8]:

```
42
```

In [10]:

```
#Pregunta 1_c: encontando el numero de cadenas comerciales monitoreadas

df.select("cadenaComercial").distinct().count()
```

Out[10]:

```
706
```

In [18]:

x

```
conteo_productos_estado = df.groupBy("estado","producto").count()
```

In [19]:

x

```
conteo_productos_estado.show()
```

```
+--------------------+--------------------+-----+
|              estado|            producto|count|
+--------------------+--------------------+-----+
|              MÉXICO|TINTE PARA EL CAB...|44007|
|              MÉXICO|         TELEVISORES|29702|
|              MÉXICO|              ACELGA| 7691|
|              MÉXICO|       QUESO. COTIJA| 4414|
|    DISTRITO FEDERAL|              AZUCAR|18078|
|              MÉXICO|      DESENFRIOL-ITO|  642|
|             JALISCO|               ARROZ|11735|
|              OAXACA|PEDIALYTE. ELECTR...|  302|
|            TLAXCALA|        AGUA SIN GAS|14505|
|VERACRUZ DE IGNAC...|              TOMATE|  652|
| MICHOACÁN DE OCAMPO|         PAN DE CAJA|13003|
|             YUCATÁN|       FLAGENASE 400|  313|
| MICHOACÁN DE OCAMPO|              ECTIVA|   39|
|             YUCATÁN|        SALSA CATSUP| 6549|
|             YUCATÁN|            CLAVULIN|  183|
|             YUCATÁN|            CAPOTENA|  271|
|             JALISCO|       FLAGENASE 400|  699|
|             HIDALGO|              VERMOX|  121|
|              OAXACA|       MAIZ POZOLERO| 1387|
|              OAXACA|                 AJO|  783|
+--------------------+--------------------+-----+
only showing top 20 rows
```

In [29]:

```
#Pregunta 1_e: Encontrando los 2 productos mas moniutoreados en cada entidad

window = Window.partitionBy(conteo_productos_estado['estado']).orderBy(conteo_productos_estado['count'].desc())

conteo_productos_estado.select('*', rank().over(window).alias('rank')).filter(col('rank') <= 2).show()
```

```
+--------------------+--------------------+-----+----+
|              estado|            producto|count|rank|
+--------------------+--------------------+-----+----+
|        QUINTANA ROO|                 FUD|34846|   1|
|        QUINTANA ROO|            REFRESCO|34367|   2|
|          NUEVO LEÓN|   DETERGENTE P/ROPA|50307|   1|
|          NUEVO LEÓN|            REFRESCO|49592|   2|
|             SINALOA|            REFRESCO|33115|   1|
|             SINALOA|   DETERGENTE P/ROPA|27177|   2|
|             TABASCO|            REFRESCO|28754|   1|
|             TABASCO|   DETERGENTE P/ROPA|26431|   2|
|     BAJA CALIFORNIA|            REFRESCO|37243|   1|
|     BAJA CALIFORNIA|   DETERGENTE P/ROPA|23395|   2|
|            TLAXCALA|            REFRESCO|43904|   1|
|            TLAXCALA|   DETERGENTE P/ROPA|41398|   2|
|COAHUILA DE ZARAGOZA|                 FUD|28613|   1|
|COAHUILA DE ZARAGOZA|            REFRESCO|26889|   2|
|                null|LECHE ULTRAPASTEU...|  804|   1|
|                null|            REFRESCO|  553|   2|
|       ESQ. SUR 125"|  PAN BLANCO BOLILLO|  130|   1|
|       ESQ. SUR 125"|    TORTILLA DE MAIZ|    2|   2|
| COL. EDUARDO GUERRA|            REFRESCO|  275|   1|
| COL. EDUARDO GUERRA|    JABON DE TOCADOR|  270|   2|
+--------------------+--------------------+-----+----+
only showing top 20 rows
```

In [47]:

x

```
df.createOrReplaceTempView("all_data")
```

In [67]:

```
sqlDF = spark.sql("SELECT cadenaComercial, producto, COUNT(*) AS numberProduct FROM all_data GROUP BY cadenaComercial, producto ORDER BY numberProduct desc")

sqlDF.show()
```

```
+--------------------+--------------------+-------------+
|     cadenaComercial|            producto|numberProduct|
+--------------------+--------------------+-------------+
|TORTILLERIAS TRAD...|    TORTILLA DE MAIZ|       206950|
|            WAL-MART|            REFRESCO|       182066|
|      BODEGA AURRERA|            REFRESCO|       173538|
|      BODEGA AURRERA|                 FUD|       136876|
|            WAL-MART|   DETERGENTE P/ROPA|       134237|
|            WAL-MART|                 FUD|       129023|
|             SORIANA|            REFRESCO|       128758|
|             SORIANA|                 FUD|       120610|
|            WAL-MART|LECHE ULTRAPASTEU...|       118766|
|             SORIANA|   DETERGENTE P/ROPA|       116610|
|      BODEGA AURRERA|LECHE ULTRAPASTEU...|       115742|
|            WAL-MART|    JABON DE TOCADOR|       107971|
|            WAL-MART|             YOGHURT|       104072|
|            WAL-MART|             CERVEZA|       102961|
|            WAL-MART|         DESODORANTE|       102042|
|            WAL-MART|             SHAMPOO|       101301|
|            WAL-MART|      CHILES EN LATA|       100236|
|MEGA COMERCIAL ME...|            REFRESCO|        98946|
|      BODEGA AURRERA|             SHAMPOO|        98942|
|            WAL-MART|            MAYONESA|        97702|
+--------------------+--------------------+-------------+
only showing top 20 rows
```

In [77]:

```
#Pregunta 1_f: encontrando la cadena comercial con mas variedad de productos

sqlDF.createOrReplaceTempView("sqlDF")

cadna_conMasProductos = spark.sql('SELECT cadenaComercial, COUNT(DISTINCT(producto)) AS numberProducts FROM sqlDF GROUP BY cadenaComercial ORDER BY numberProducts desc')

cadna_conMasProductos.show()
```

```
+--------------------+--------------+
|     cadenaComercial|numberProducts|
+--------------------+--------------+
|             SORIANA|          1059|
|            WAL-MART|          1051|
|MEGA COMERCIAL ME...|          1049|
|  COMERCIAL MEXICANA|          1036|
|            CHEDRAUI|          1026|
|     MERCADO SORIANA|          1024|
|      BODEGA AURRERA|          1012|
|HIPERMERCADO SORIANA|          1006|
|              H.E.B.|          1001|
|        SORIANA PLUS|           999|
|       SORIANA SUPER|           996|
|BODEGA COMERCIAL ...|           979|
|        I.S.S.S.T.E.|           937|
|            SUPERAMA|           936|
|              S MART|           851|
|SUPERMERCADOS SAN...|           849|
|              SUMESA|           848|
|         CITY MARKET|           844|
|FARMACIA GUADALAJARA|           819|
|            CASA LEY|           808|
+--------------------+--------------+
only showing top 20 rows
```