

```

1 #configuración en google colab de spark y pyspark
2 from google.colab import drive
3 drive.mount('/content/gdrive')

```

Mounted at /content/gdrive

```

1 #instalar java y spark
2 !apt-get install openjdk-11-jdk-headless -qq > /dev/null
3 !wget -q https://dlcdn.apache.org/spark/spark-3.5.3/spark-3.5.3-bin-hadoop3.tgz
4 !tar xf spark-3.5.3-bin-hadoop3.tgz
5 !pip install -q findspark

```

```

1 import os
2 os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
3 os.environ["SPARK_HOME"] = "/content/spark-3.5.3-bin-hadoop3"

```

```

1 import findspark
2 findspark.init()
3 from pyspark.sql import SparkSession
4 spark = SparkSession.builder.master("local[*]").getOrCreate()
5 sc = spark.sparkContext

```

```
1 spark
```

SparkSession - in-memory

SparkContext

[Spark UI](#)

Version
v3.5.3
Master
local[*]
AppName
pyspark-shell

```
1 sc
```

SparkContext

[Spark UI](#)

Version
v3.5.3
Master
local[*]
AppName
pyspark-shell

```

1 # Load csv Dataset
2 df=spark.read.csv('gdrive/MyDrive/st0263-242/bigdata/datasets/covid19/Casos_positivos_de_COVID-19_en_Colombia-100K.csv',inferSchema=True)

```

```
1 df.columns
```

['fecha reporte web',
 'ID de caso',
 'Fecha de notificación',
 'Código DIVIPOLA departamento',
 'Nombre departamento',
 'Código DIVIPOLA municipio',
 'Nombre municipio',
 'Edad',
 'Unidad de medida de edad',
 'Sexo',
 'Tipo de contagio',
 'Ubicación del caso',
 'Estado',
 'Código ISO del país',
 'Nombre del país',
 'Recuperado',
 'Fecha de inicio de síntomas',
 'Fecha de muerte',
 'Fecha de diagnóstico',
 'Fecha de recuperación',
 'Tipo de recuperación',
 'Pertenencia étnica',
 'Nombre del grupo étnico']

```
1 df.printSchema()
```

```

root
|-- fecha reporte web: string (nullable = true)
|-- ID de caso: integer (nullable = true)
|-- Fecha de notificación: string (nullable = true)
|-- Código DIVIPOLA departamento: integer (nullable = true)
|-- Nombre departamento: string (nullable = true)
|-- Código DIVIPOLA municipio: integer (nullable = true)
|-- Nombre municipio: string (nullable = true)
|-- Edad: integer (nullable = true)
|-- Unidad de medida de edad: integer (nullable = true)
|-- Sexo: string (nullable = true)
|-- Tipo de contagio: string (nullable = true)
|-- Ubicación del caso: string (nullable = true)
|-- Estado: string (nullable = true)
|-- Código ISO del país: integer (nullable = true)
|-- Nombre del país: string (nullable = true)
|-- Recuperado: string (nullable = true)
|-- Fecha de inicio de síntomas: string (nullable = true)
|-- Fecha de muerte: string (nullable = true)
|-- Fecha de diagnóstico: string (nullable = true)
|-- Fecha de recuperación: string (nullable = true)
|-- Tipo de recuperación: string (nullable = true)
|-- Pertenencia étnica: integer (nullable = true)
|-- Nombre del grupo étnico: string (nullable = true)

```

```
1 df.select('fecha reporte web', 'ID de caso', 'Sexo', 'Tipo de contagio').show(10, False)
```

```

+-----+-----+-----+-----+
|fecha reporte web|ID de caso|Sexo|Tipo de contagio|
+-----+-----+-----+-----+
|6/3/2020 0:00:00|1         |F   |Importado         |
|9/3/2020 0:00:00|2         |M   |Importado         |
|9/3/2020 0:00:00|3         |F   |Importado         |
|11/3/2020 0:00:00|4        |M   |Relacionado       |
|11/3/2020 0:00:00|5        |M   |Relacionado       |
|11/3/2020 0:00:00|6        |F   |Relacionado       |
|11/3/2020 0:00:00|7        |F   |Importado         |
|11/3/2020 0:00:00|8        |F   |Importado         |
|11/3/2020 0:00:00|9        |F   |Importado         |
|12/3/2020 0:00:00|10       |F   |Importado         |
+-----+-----+-----+-----+
only showing top 10 rows

```

```
1 df=df.withColumnRenamed('fecha reporte web', 'fecha_reporte')
```

```
1 df=df.withColumnsRenamed({'ID de caso': 'ID', 'Nombre del país': 'pais', 'Nombre departamento': 'departamento', 'Nombre municipio':
```

```
1 df=df.withColumn('Edad_mas_10', df['Edad'] + 10)
2 df.show(5, False)
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|fecha_reporte  |ID |Fecha de notificación|Código DIVIPOLA departamento|departamento|Código DIVIPOLA municipio|municipio|Edad|Unic
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|6/3/2020 0:00:00|1  |2/3/2020 0:00:00    |11                          |BOGOTA      |11001                    |BOGOTA  |19 |1
|9/3/2020 0:00:00|2  |6/3/2020 0:00:00    |76                          |VALLE       |76111                    |BUGA    |34 |1
|9/3/2020 0:00:00|3  |7/3/2020 0:00:00    |5                           |ANTIOQUIA   |5001                     |MEDELLIN|50 |1
|11/3/2020 0:00:00|4  |9/3/2020 0:00:00    |5                           |ANTIOQUIA   |5001                     |MEDELLIN|55 |1
|11/3/2020 0:00:00|5  |9/3/2020 0:00:00    |5                           |ANTIOQUIA   |5001                     |MEDELLIN|25 |1
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows

```

Double-click (or enter) to edit

```
1 columns_to_remove=['Nombre del grupo étnico', 'Pertenencia étnica']
2 df=df.drop(*columns_to_remove)
```

```
1 df.printSchema()
```

```

root
|-- fecha_reporte: string (nullable = true)
|-- ID: integer (nullable = true)
|-- Fecha de notificación: string (nullable = true)
|-- Código DIVIPOLA departamento: integer (nullable = true)
|-- departamento: string (nullable = true)
|-- Código DIVIPOLA municipio: integer (nullable = true)
|-- municipio: string (nullable = true)
|-- Edad: integer (nullable = true)

```

```
-- Unidad de medida de edad: integer (nullable = true)
-- Sexo: string (nullable = true)
-- Tipo de contagio: string (nullable = true)
-- Ubicación: string (nullable = true)
-- Estado: string (nullable = true)
-- Código ISO del país: integer (nullable = true)
-- pais: string (nullable = true)
-- Recuperado: string (nullable = true)
-- Fecha de inicio de síntomas: string (nullable = true)
-- Fecha de muerte: string (nullable = true)
-- Fecha de diagnóstico: string (nullable = true)
-- Fecha de recuperación: string (nullable = true)
-- Tipo de recuperación: string (nullable = true)
-- age_group: string (nullable = true)
-- Edad_mas_10: integer (nullable = true)
```

```
1 filtered_df=df.filter(df.Sexo == 'F')
2 filtered_df.show(5, False)
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| fecha_reporte | ID | Fecha de notificación | Código DIVIPOLA | departamento | departamento | Código DIVIPOLA | municipio | municipio | Edad | Unic |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 6/3/2020 0:00:00 | 1 | 2/3/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | 19 | 1 |
| 9/3/2020 0:00:00 | 3 | 7/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 50 | 1 |
| 11/3/2020 0:00:00 | 6 | 10/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5360 | ITAGUI | ITAGUI | 27 | 1 |
| 11/3/2020 0:00:00 | 7 | 8/3/2020 0:00:00 | 13001 | CARTAGENA | CARTAGENA | 13001 | CARTAGENA | CARTAGENA | 85 | 1 |
| 11/3/2020 0:00:00 | 8 | 9/3/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | 22 | 1 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows

```

```
1 filtered2_df=df.filter("Sexo=='F' and Edad>20 and Recuperado like '%Recuperado%' and municipio like 'MEDELLIN'")
2 filtered2_df.show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| fecha_reporte | ID | Fecha de notificación | Código DIVIPOLA | departamento | departamento | Código DIVIPOLA | municipio | municipio | Edad | Unic |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 9/3/2020 0:00:00 | 3 | 7/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 50 | 1 |
| 14/3/2020 0:00:00 | 20 | 11/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 26 | 1 |
| 19/3/2020 0:00:00 | 108 | 17/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 57 | 1 |
| 20/3/2020 0:00:00 | 131 | 15/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 22 | 1 |
| 20/3/2020 0:00:00 | 135 | 17/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 44 | 1 |
| 20/3/2020 0:00:00 | 141 | 17/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 62 | 1 |
| 20/3/2020 0:00:00 | 142 | 20/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 35 | 1 |
| 22/3/2020 0:00:00 | 238 | 16/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 61 | 1 |
| 23/3/2020 0:00:00 | 268 | 19/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 37 | 1 |
| 23/3/2020 0:00:00 | 271 | 18/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 21 | 1 |
| 23/3/2020 0:00:00 | 272 | 18/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 47 | 1 |
| 23/3/2020 0:00:00 | 275 | 23/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 21 | 1 |
| 23/3/2020 0:00:00 | 292 | 19/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 43 | 1 |
| 23/3/2020 0:00:00 | 293 | 18/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 30 | 1 |
| 23/3/2020 0:00:00 | 294 | 21/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 30 | 1 |
| 23/3/2020 0:00:00 | 296 | 19/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 25 | 1 |
| 25/3/2020 0:00:00 | 438 | 20/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 36 | 1 |
| 25/3/2020 0:00:00 | 450 | 19/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 24 | 1 |
| 25/3/2020 0:00:00 | 466 | 19/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 32 | 1 |
| 28/3/2020 0:00:00 | 595 | 16/3/2020 0:00:00 | 5 | ANTIOQUIA | ANTIOQUIA | 5001 | MEDELLIN | MEDELLIN | 27 | 1 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

```

```
1 from pyspark.sql.functions import udf
2 from pyspark.sql.types import StringType, DoubleType, IntegerType
```

```
1 age_udf=udf(lambda age: "kid" if age < 18 else ('young adult' if age<=30 else 'senior'), StringType())
2 df=df.withColumn("age_group", age_udf(df.Edad))
3 df.filter(df['age_group']=='young adult').orderBy('Edad', ascending=False).show(10, False)
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| fecha_reporte | ID de caso | Fecha de notificación | Código DIVIPOLA | departamento | departamento | Código DIVIPOLA | municipio | municipio | Nombre |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 19/6/2020 0:00:00 | 62406 | 10/6/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | BOGOTA |
| 21/3/2020 0:00:00 | 192 | 20/3/2020 0:00:00 | 76 | VALLE | VALLE | 76001 | CALI | CALI | CALI |
| 19/6/2020 0:00:00 | 62398 | 17/6/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | BOGOTA |
| 27/3/2020 0:00:00 | 517 | 23/3/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | BOGOTA |
| 19/6/2020 0:00:00 | 62380 | 10/6/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | BOGOTA |
| 22/3/2020 0:00:00 | 213 | 19/3/2020 0:00:00 | 20 | CESAR | CESAR | 20001 | VALLERÍA | VALLERÍA | VALLERÍA |
| 19/6/2020 0:00:00 | 62617 | 16/6/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | BOGOTA |
| 13/3/2020 0:00:00 | 15 | 13/3/2020 0:00:00 | 50 | META | META | 50001 | VILLAVIEJA | VILLAVIEJA | VILLAVIEJA |
| 19/6/2020 0:00:00 | 62384 | 7/6/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | BOGOTA |
| 23/3/2020 0:00:00 | 248 | 23/3/2020 0:00:00 | 11 | BOGOTA | BOGOTA | 11001 | BOGOTA | BOGOTA | BOGOTA |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 10 rows

```

```
1 df.groupBy('departamento').count().orderBy('count', ascending=False).show(10, False)
```

```

+-----+-----+
|departamento|count|
+-----+-----+
|BOGOTA       |30016|
|BARRANQUILLA|13065|
|ATLANTICO    |10994|
|VALLE        |10404|
|CARTAGENA    |8333 |
|ANTIOQUIA    |4554 |
|NARIÑO       |3520 |
|CUNDINAMARCA|2827 |
|AMAZONAS     |2317 |
|CHOCO        |1636 |
+-----+-----+
only showing top 10 rows

```

```
1 df.groupBy('municipio').count().orderBy('count', ascending=False).show(10, False)
```

```

+-----+-----+
|municipio   |count|
+-----+-----+
|BOGOTA       |30016|
|BARRANQUILLA|13065|
|CARTAGENA    |8333 |
|CALI         |7747 |
|SOLEDAD      |6233 |
|LETICIA      |2194 |
|MEDELLIN     |2137 |
|TUMACO       |1501 |
|BUENAVENTURA|1453 |
|QUIBDO       |1367 |
+-----+-----+
only showing top 10 rows

```

```
1 df.dropna(subset=['Fecha de inicio de síntomas']).groupBy('Fecha de inicio de síntomas').count().orderBy('count', ascending=False).show(10, False)
```

```

+-----+-----+
|Fecha de inicio de síntomas|count|
+-----+-----+
|10/6/2020 0:00:00          |2731 |
|16/6/2020 0:00:00          |2558 |
|18/6/2020 0:00:00          |2479 |
|12/6/2020 0:00:00          |2452 |
|1/6/2020 0:00:00           |2429 |
|8/6/2020 0:00:00           |2390 |
|17/6/2020 0:00:00          |2344 |
|5/6/2020 0:00:00           |2266 |
|9/6/2020 0:00:00           |2224 |
|19/6/2020 0:00:00          |2162 |
+-----+-----+
only showing top 10 rows

```

```
1 df.select('edad', 'age_group', 'ID').groupBy('edad', 'age_group').count().orderBy('count', ascending=False).show(10, False)
```

```

+-----+-----+-----+
|edad|  age_group|count|
+-----+-----+-----+
| 30|young adult| 2735|
| 29|young adult| 2611|
| 31|   senior| 2569|
| 28|young adult| 2540|
| 27|young adult| 2494|
| 26|young adult| 2436|
| 33|   senior| 2371|
| 32|   senior| 2362|
| 25|young adult| 2335|
| 34|   senior| 2310|
| 35|   senior| 2292|
| 24|young adult| 2214|
| 36|   senior| 2175|
| 37|   senior| 2132|
| 38|   senior| 2098|
| 40|   senior| 2050|
| 23|young adult| 2041|
| 39|   senior| 1985|
| 22|young adult| 1879|
| 41|   senior| 1783|
+-----+-----+-----+

```

only showing top 20 rows

```
1 from pyspark.sql.functions import col
2 df.filter(col('Fecha de muerte').isNotNull()).groupBy('age_group').count().orderBy('count', ascending=False).show()
```

```

+-----+-----+
| age_group|count|
+-----+-----+
|    senior| 5435|
| young adult| 151|
|      kid| 47|
+-----+-----+

```

```
1 df.filter(col('Fecha de muerte').isNotNull()).show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| fecha_reporte| ID| Fecha de notificación| Código DIVIPOLA| departamento| departamento| Código DIVIPOLA| municipio| municipio| Edad|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 20/3/2020 0:00:00| 152| 18/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 65
| 20/3/2020 0:00:00| 153| 18/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 53
| 20/3/2020 0:00:00| 157| 20/3/2020 0:00:00| 47001| STA MARTA D.E.| 47001| SANTA MARTA| 88
| 21/3/2020 0:00:00| 188| 17/3/2020 0:00:00| 76| VALLE| 76892| YUMBO| 76
| 16/3/2020 0:00:00| 197| 13/3/2020 0:00:00| 13001| CARTAGENA| 13001| CARTAGENA| 58
| 22/3/2020 0:00:00| 232| 20/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 59
| 23/3/2020 0:00:00| 250| 23/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 76
| 23/3/2020 0:00:00| 261| 23/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 67
| 23/3/2020 0:00:00| 282| 20/3/2020 0:00:00| 76| VALLE| 76001| CALI| 46
| 23/3/2020 0:00:00| 286| 21/3/2020 0:00:00| 41| HUILA| 41001| NEIVA| 83
| 24/3/2020 0:00:00| 314| 20/3/2020 0:00:00| 52| NARIÑO| 52356| IPIALES| 67
| 24/3/2020 0:00:00| 326| 21/3/2020 0:00:00| 66| RISARALDA| 66001| PEREIRA| 67
| 24/3/2020 0:00:00| 389| 23/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 73
| 24/3/2020 0:00:00| 415| 21/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 67
| 25/3/2020 0:00:00| 425| 21/3/2020 0:00:00| 76| VALLE| 76834| TULUA| 86
| 26/3/2020 0:00:00| 491| 25/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 84
| 27/3/2020 0:00:00| 510| 23/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 77
| 27/3/2020 0:00:00| 514| 22/3/2020 0:00:00| 11| BOGOTA| 11001| BOGOTA| 77
| 27/3/2020 0:00:00| 533| 24/3/2020 0:00:00| 8001| BARRANQUILLA| 8001| BARRANQUILLA| 25
| 27/3/2020 0:00:00| 535| 24/3/2020 0:00:00| 8| ATLANTICO| 8758| SOLEDAD| 56
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

only showing top 20 rows

Using SparkSQL

```
1 df.createOrReplaceTempView("covid19")
```

```
1 spark.sql("select departamento, count(*) as count from covid19 group by departamento order by count desc").show(10)
```

```

+-----+-----+
| departamento|count|
+-----+-----+
| BOGOTA| 30016|
| BARRANQUILLA| 13065|
| ATLANTICO| 10994|
| VALLE| 10404|
| CARTAGENA| 8333|
| ANTIOQUIA| 4554|
| NARIÑO| 3520|
| CUNDINAMARCA| 2827|
| AMAZONAS| 2317|
| CHOCO| 1636|
+-----+-----+
only showing top 10 rows

```

```
1 spark.sql("select municipio, count(*) as count from covid19 group by municipio order by count desc").show(10)
```

```

+-----+-----+
| municipio|count|
+-----+-----+
| BOGOTA| 30016|
| BARRANQUILLA| 13065|
| CARTAGENA| 8333|
| CALI| 7747|
| SOLEDAD| 6233|
| LETICIA| 2194|
| MEDELLIN| 2137|
| TUMACO| 1501|
| BUENAVENTURA| 1453|
| QUIBDO| 1367|
+-----+-----+

```

only showing top 10 rows

```
1 spark.sql("""select `Fecha de inicio de síntomas` as fecha, count(*) as count from covid19 where
2 `Fecha de inicio de síntomas` is not NULL group by fecha order by count desc limit 10""").show()
```

```
↗ +-----+-----+
|          fecha|count|
+-----+-----+
|10/6/2020 0:00:00| 2731|
|16/6/2020 0:00:00| 2558|
|18/6/2020 0:00:00| 2479|
|12/6/2020 0:00:00| 2452|
| 1/6/2020 0:00:00| 2429|
| 8/6/2020 0:00:00| 2390|
|17/6/2020 0:00:00| 2344|
| 5/6/2020 0:00:00| 2266|
| 9/6/2020 0:00:00| 2224|
|19/6/2020 0:00:00| 2162|
+-----+-----+
```

```
1 spark.sql('select edad, age_group,count(*) as count from covid19 group by edad, age_group order by count desc').show()
```

```
↗ +-----+-----+
|edad|  age_group|count|
+-----+-----+
| 30|young adult| 2735|
| 29|young adult| 2611|
| 31|  senior| 2569|
| 28|young adult| 2540|
| 27|young adult| 2494|
| 26|young adult| 2436|
| 33|  senior| 2371|
| 32|  senior| 2362|
| 25|young adult| 2335|
| 34|  senior| 2310|
| 35|  senior| 2292|
| 24|young adult| 2214|
| 36|  senior| 2175|
| 37|  senior| 2132|
| 38|  senior| 2098|
| 40|  senior| 2050|
| 23|young adult| 2041|
| 39|  senior| 1985|
| 22|young adult| 1879|
| 41|  senior| 1783|
+-----+-----+
only showing top 20 rows
```

```
1 spark.sql('''select age_group, count(*) as count from covid19 where
2 `Fecha de muerte` is not null group by age_group order by count desc''').show()
```

```
↗ +-----+-----+
|  age_group|count|
+-----+-----+
|  senior| 5435|
|young adult| 147|
|  kid| 51|
+-----+-----+
```