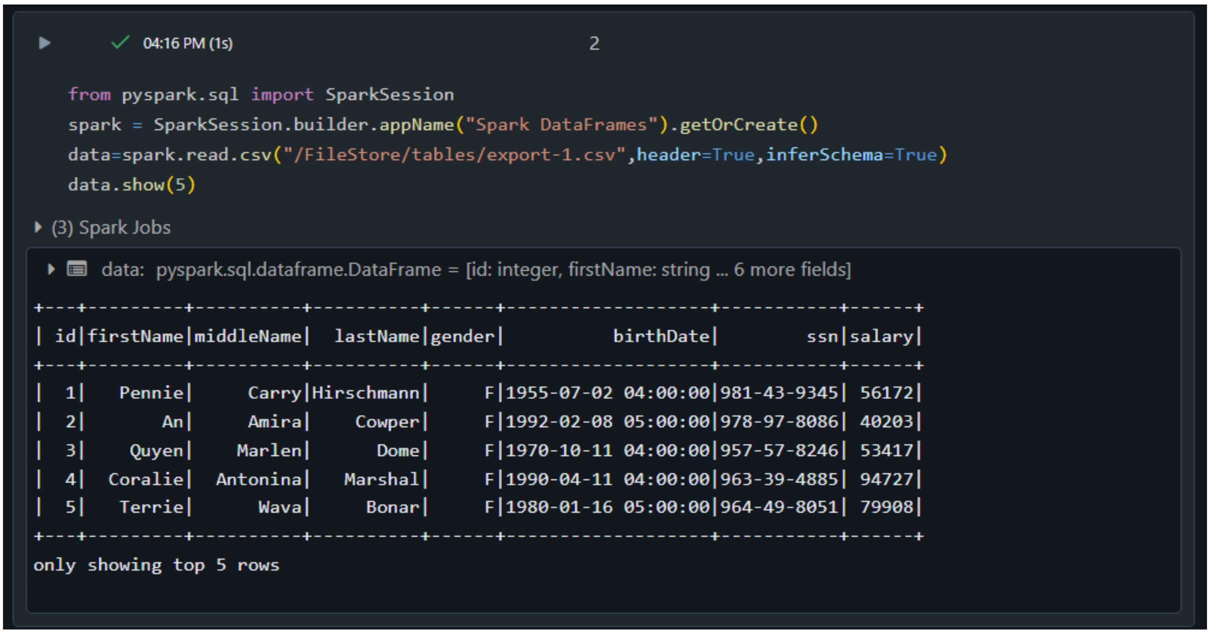
## Ads Day2 Assignment

Reading as csv files



▶ ✓ 04:16 PM (1s) 2

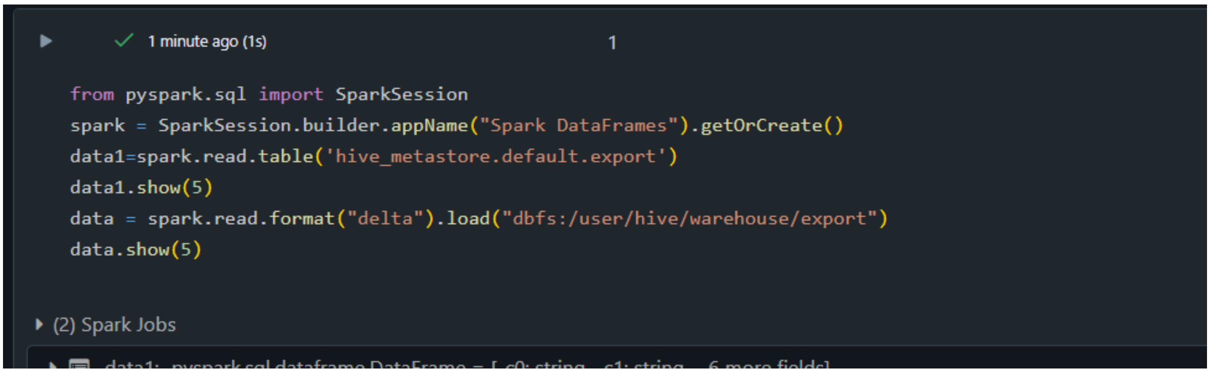
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Spark DataFrames").getOrCreate()
data=spark.read.csv("/FileStore/tables/export-1.csv",header=True,inferSchema=True)
data.show(5)
```

▸ (3) Spark Jobs

▸ data: pyspark.sql.dataframe.DataFrame = [id: integer, firstName: string ... 6 more fields]

```
+---+---------+----------+----------+------+-------------------+-----------+------+
| id|firstName|middleName|  lastName|gender|          birthDate|        ssn|salary|
+---+---------+----------+----------+------+-------------------+-----------+------+
|  1|   Pennie|     Carry|Hirschmann|     F|1955-07-02 04:00:00|981-43-9345| 56172|
|  2|       An|     Amira|    Cowper|     F|1992-02-08 05:00:00|978-97-8086| 40203|
|  3|    Quyen|    Marlen|      Dome|     F|1970-10-11 04:00:00|957-57-8246| 53417|
|  4|  Coralie|  Antonina|   Marshal|     F|1990-04-11 04:00:00|963-39-4885| 94727|
|  5|   Terrie|      Wava|     Bonar|     F|1980-01-16 05:00:00|964-49-8051| 79908|
+---+---------+----------+----------+------+-------------------+-----------+------+
only showing top 5 rows
```

Delta tables

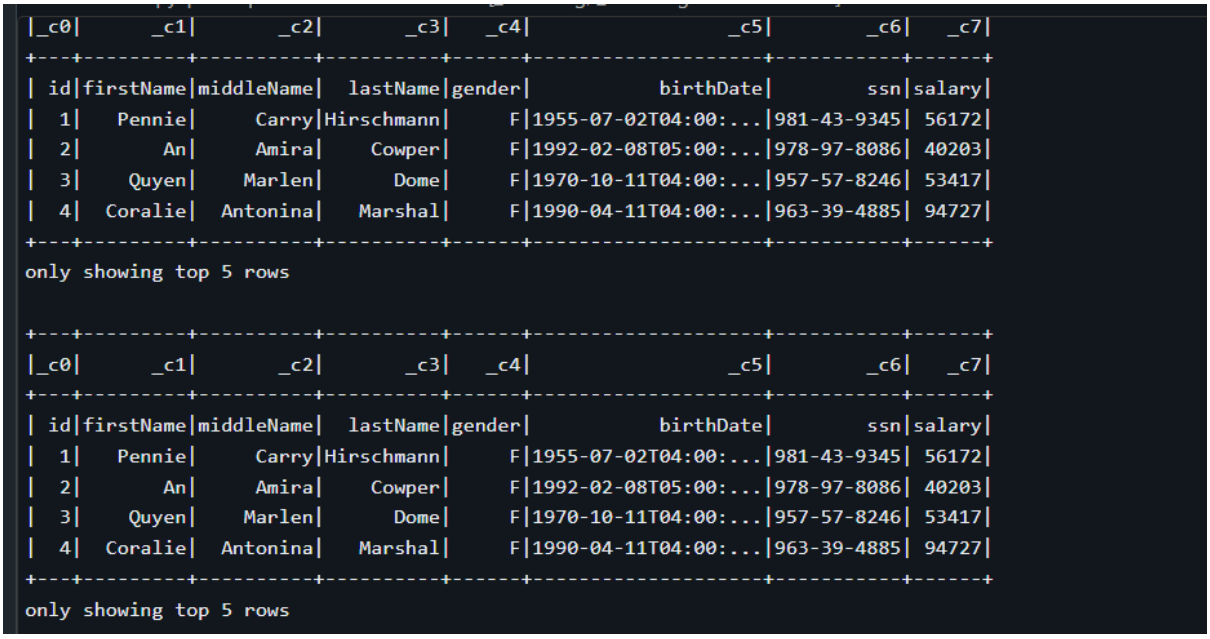


▶ ✓ 1 minute ago (1s) 1

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Spark DataFrames").getOrCreate()
data1=spark.read.table('hive_metastore.default.export')
data1.show(5)
data = spark.read.format("delta").load("dbfs:/user/hive/warehouse/export")
data.show(5)
```

▸ (2) Spark Jobs

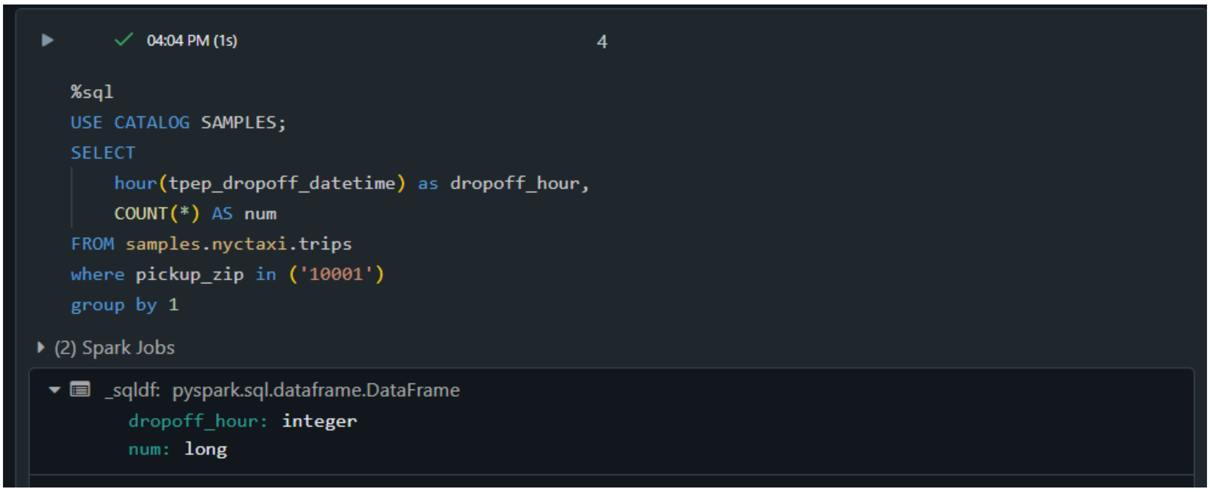
▸ data1: pyspark.sql.dataframe.DataFrame = [_c0: string, _c1: string ... 6 more fields]



```
|_c0|      _c1|       _c2|       _c3|   _c4|                 _c5|        _c6|   _c7|
+---+---------+----------+----------+------+--------------------+-----------+------+
| id|firstName|middleName|  lastName|gender|           birthDate|        ssn|salary|
|  1|   Pennie|     Carry|Hirschmann|     F|1955-07-02T04:00:...|981-43-9345| 56172|
|  2|       An|     Amira|    Cowper|     F|1992-02-08T05:00:...|978-97-8086| 40203|
|  3|    Quyen|    Marlen|      Dome|     F|1970-10-11T04:00:...|957-57-8246| 53417|
|  4|  Coralie|  Antonina|   Marshal|     F|1990-04-11T04:00:...|963-39-4885| 94727|
+---+---------+----------+----------+------+--------------------+-----------+------+
only showing top 5 rows

+---+---------+----------+----------+------+--------------------+-----------+------+
|_c0|      _c1|       _c2|       _c3|   _c4|                 _c5|        _c6|   _c7|
+---+---------+----------+----------+------+--------------------+-----------+------+
| id|firstName|middleName|  lastName|gender|           birthDate|        ssn|salary|
|  1|   Pennie|     Carry|Hirschmann|     F|1955-07-02T04:00:...|981-43-9345| 56172|
|  2|       An|     Amira|    Cowper|     F|1992-02-08T05:00:...|978-97-8086| 40203|
|  3|    Quyen|    Marlen|      Dome|     F|1970-10-11T04:00:...|957-57-8246| 53417|
|  4|  Coralie|  Antonina|   Marshal|     F|1990-04-11T04:00:...|963-39-4885| 94727|
+---+---------+----------+----------+------+--------------------+-----------+------+
only showing top 5 rows
```

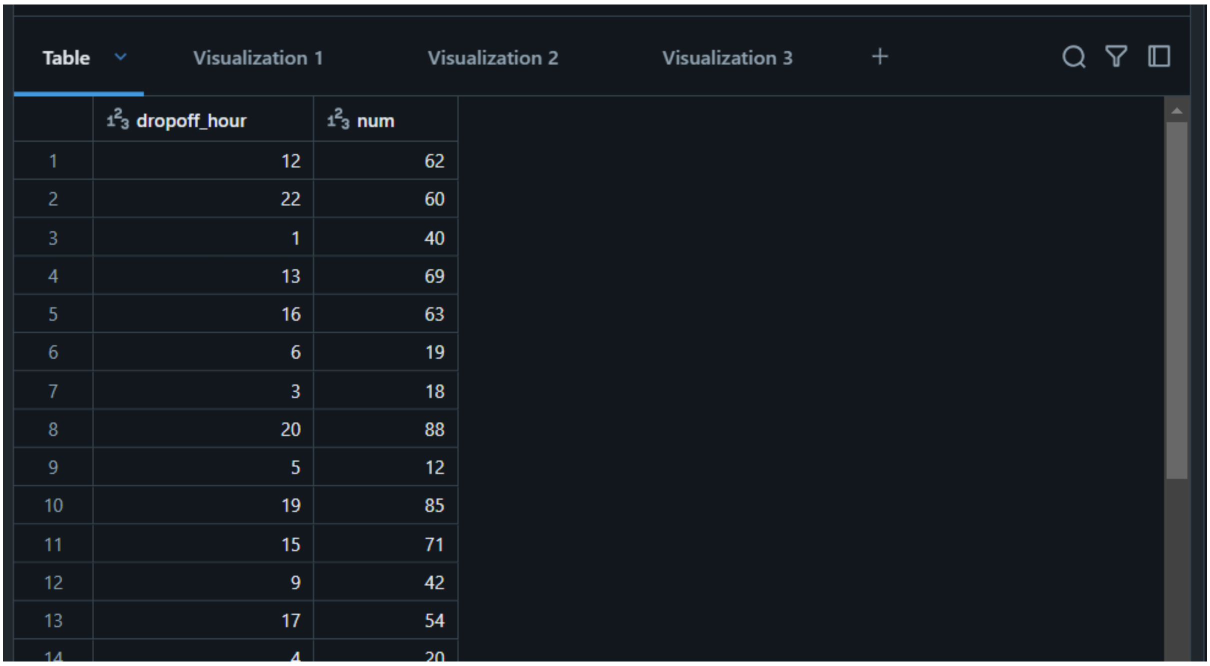
Data visualization



```
%sql
USE CATALOG SAMPLES;
SELECT
    hour(tpep_dropoff_datetime) as dropoff_hour,
    COUNT(*) AS num
FROM samples.nyctaxi.trips
where pickup_zip in ('10001')
group by 1
```

(2) Spark Jobs

```
_sqldf: pyspark.sql.dataframe.DataFrame
    dropoff_hour: integer
    num: long
```



| | dropoff_hour | num |
|---|---|---|
| 1 | 12 | 62 |
| 2 | 22 | 60 |
| 3 | 1 | 40 |
| 4 | 13 | 69 |
| 5 | 16 | 63 |
| 6 | 6 | 19 |
| 7 | 3 | 18 |
| 8 | 20 | 88 |
| 9 | 5 | 12 |
| 10 | 19 | 85 |
| 11 | 15 | 71 |
| 12 | 9 | 42 |
| 13 | 17 | 54 |
| 14 | 4 | 20 |

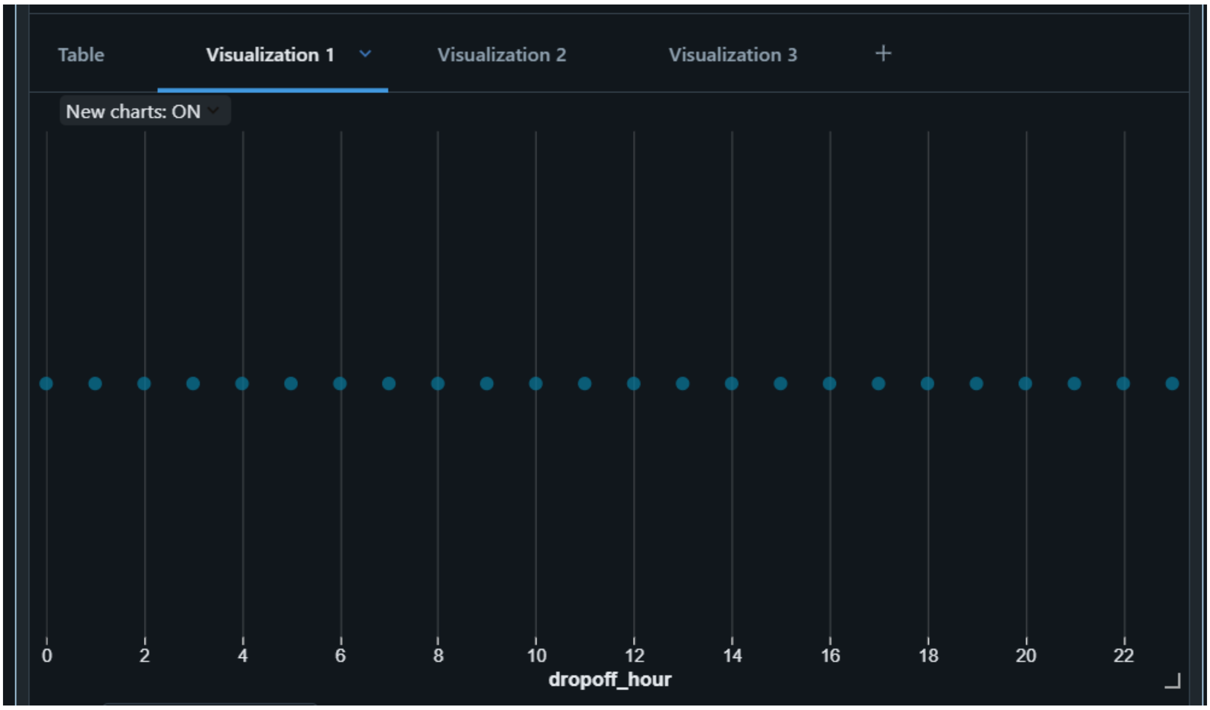


Table
Visualization 1
Visualization 2
Visualization 3
+
New charts: ON
0
2
4
6
8
10
12
14
16
18
20
22
dropoff_hour

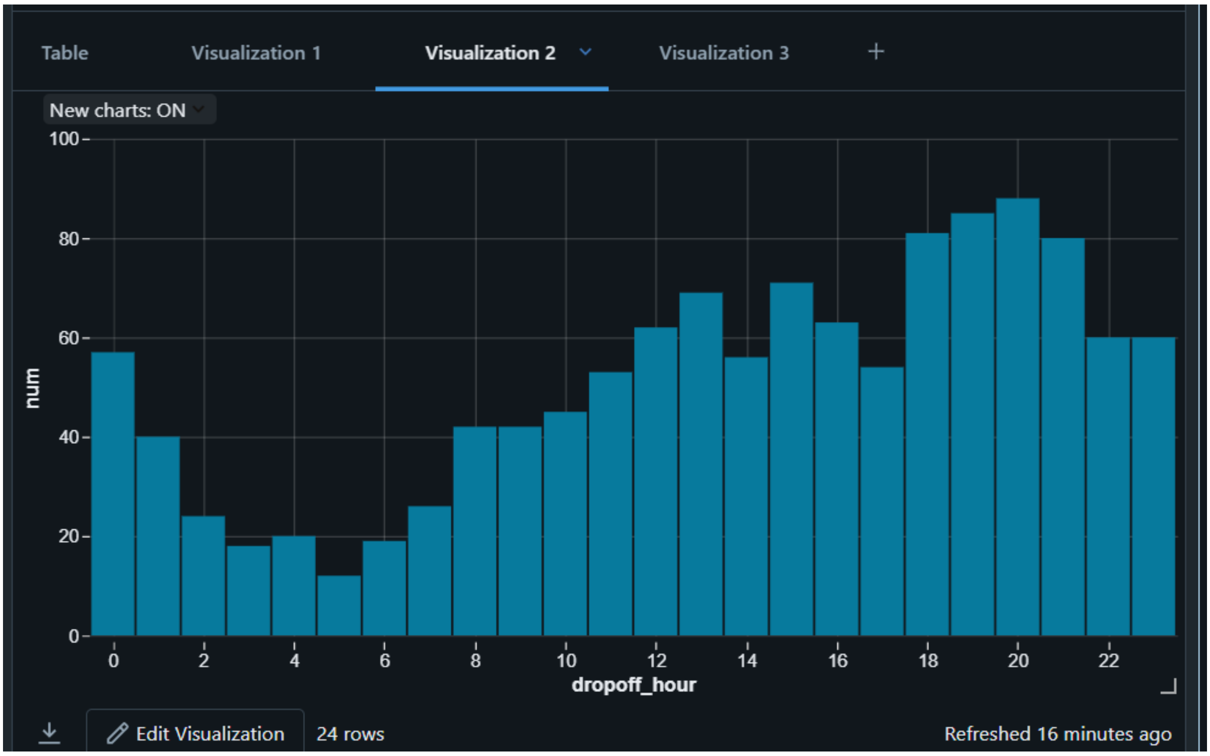
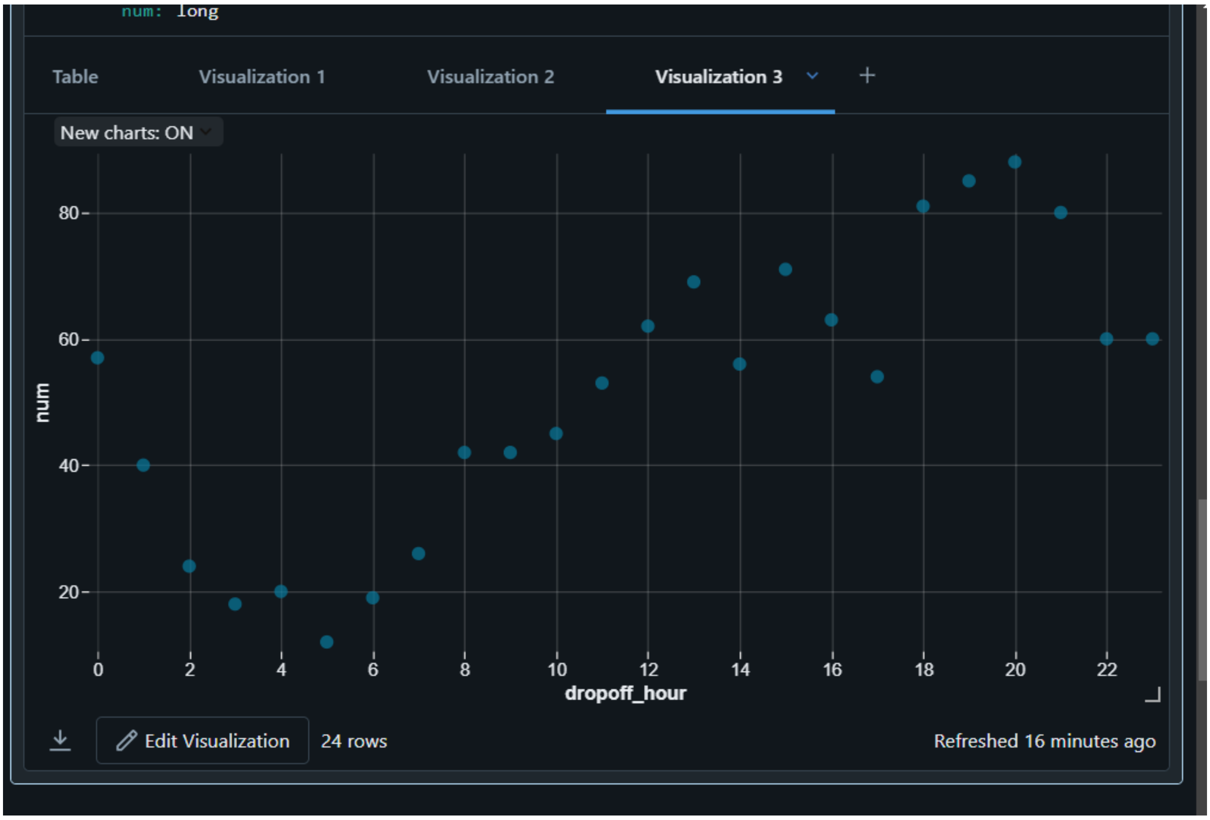
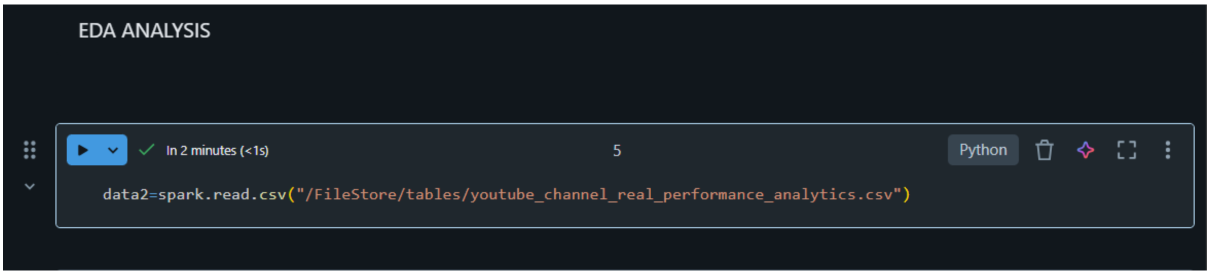


Table
Visualization 1
Visualization 2
Visualization 3
+
New charts: ON
100
80
60
40
20
0
num
0
2
4
6
8
10
12
14
16
18
20
22
dropoff_hour
Edit Visualization
24 rows
Refreshed 16 minutes ago



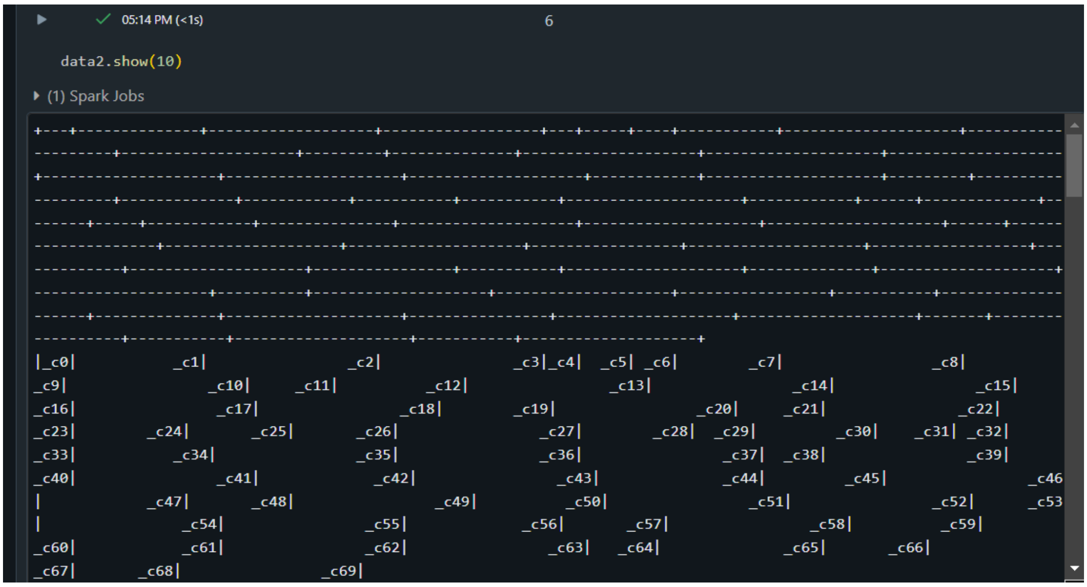
```
num: long
```

EDA Analysis



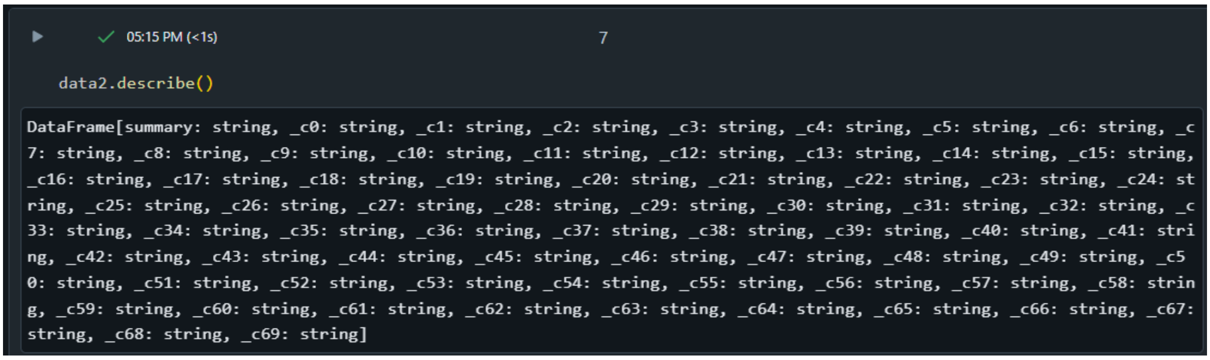
EDA ANALYSIS

```
data2=spark.read.csv("/FileStore/tables/youtube_channel_real_performance_analytics.csv")
```



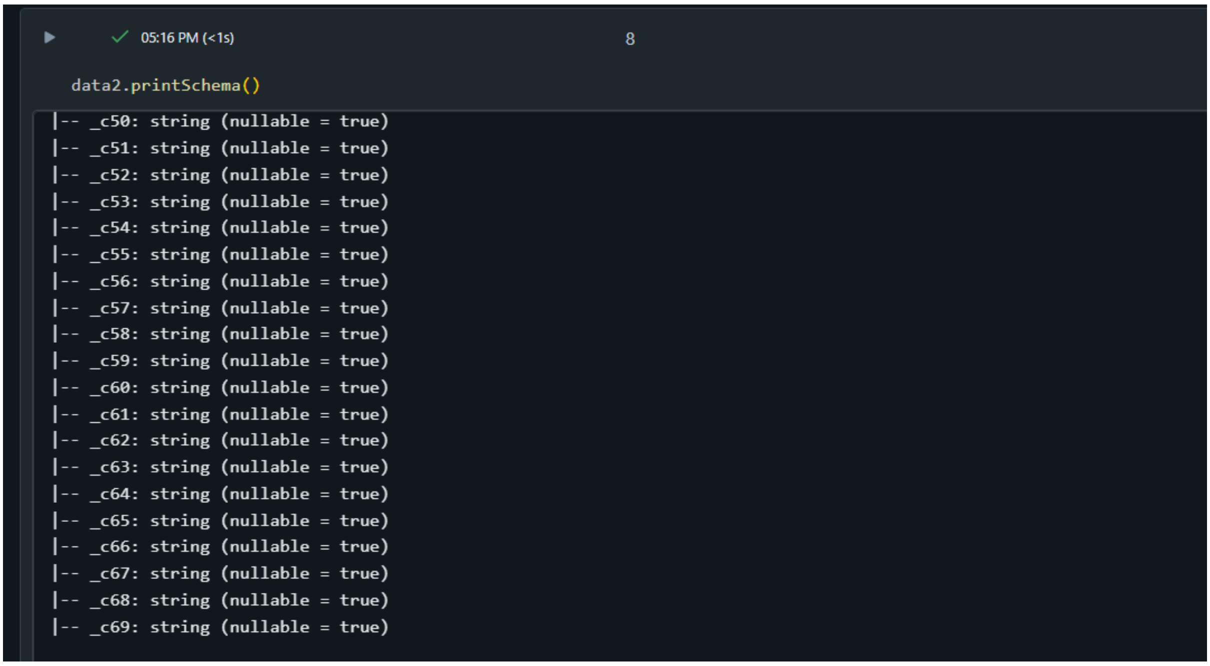
```
data2.show(10)
```

```
+---+-------------+------------------+-----------------+---+-----+----+----------+-------------------+----------
---------+-------------------+--------+--------------+------------------+-------------------+-------------------
+-------------------+-------------------+-------------------+-----------+-------------------+---------+---------
---------+------------+-----------+-----------+----------+-------------------+------------+-----+-------------+-
------+-----+-----------+--------------+-------------------+-------------------+-------------------+------+-----
--------------+-------------------+------------------+-----------------+-------------------+-----------------+--
----------+-------------------+---------------+----------+-------------------+--------------+------------------+
-------------------+----------+-------------------+-------------------+----------------+----------+-------------
------+-------------+-------------------+---------------+-------------------+-------------------+-------+-------
----------+----------+-------------------+-----------+------------------+
|_c0|          _c1|               _c2|              _c3|_c4|  _c5| _c6|       _c7|                _c8|
_c9|               _c10|     _c11|         _c12|               _c13|               _c14|               _c15|
_c16|               _c17|               _c18|       _c19|               _c20|     _c21|              _c22|
_c23|       _c24|       _c25|      _c26|               _c27|        _c28| _c29|        _c30|    _c31| _c32|
_c33|          _c34|               _c35|               _c36|               _c37|  _c38|               _c39|
_c40|               _c41|            _c42|               _c43|             _c44|        _c45|               _c46
|           _c47|       _c48|               _c49|         _c50|               _c51|               _c52|     _c53
|               _c54|               _c55|            _c56|       _c57|               _c58|         _c59|
_c60|           _c61|               _c62|               _c63|   _c64|             _c65|      _c66|
_c67|      _c68|               _c69|
```



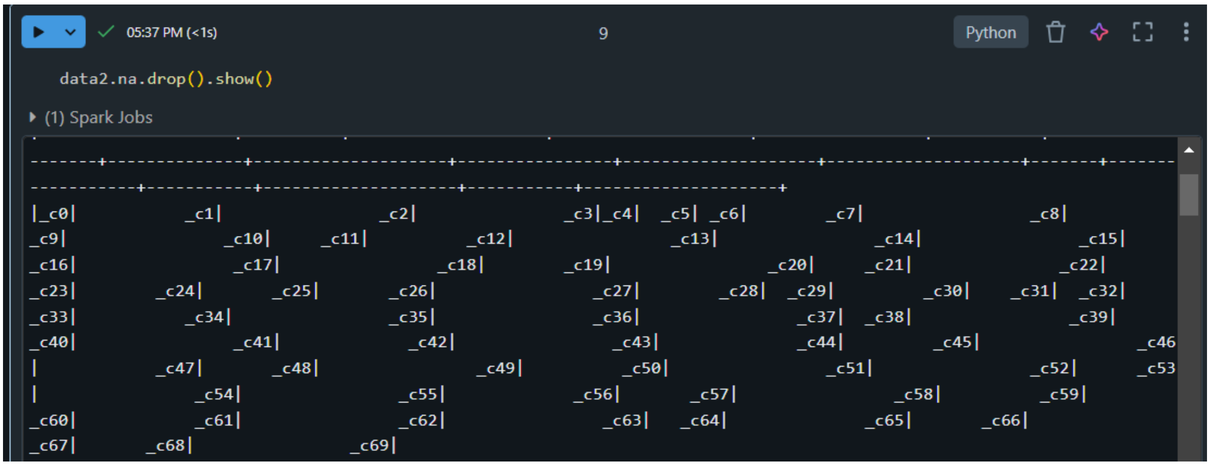
```
data2.describe()
```

```
DataFrame[summary: string, _c0: string, _c1: string, _c2: string, _c3: string, _c4: string, _c5: string, _c6: string, _c
7: string, _c8: string, _c9: string, _c10: string, _c11: string, _c12: string, _c13: string, _c14: string, _c15: string,
_c16: string, _c17: string, _c18: string, _c19: string, _c20: string, _c21: string, _c22: string, _c23: string, _c24: st
ring, _c25: string, _c26: string, _c27: string, _c28: string, _c29: string, _c30: string, _c31: string, _c32: string, _c
33: string, _c34: string, _c35: string, _c36: string, _c37: string, _c38: string, _c39: string, _c40: string, _c41: stri
ng, _c42: string, _c43: string, _c44: string, _c45: string, _c46: string, _c47: string, _c48: string, _c49: string, _c5
0: string, _c51: string, _c52: string, _c53: string, _c54: string, _c55: string, _c56: string, _c57: string, _c58: strin
g, _c59: string, _c60: string, _c61: string, _c62: string, _c63: string, _c64: string, _c65: string, _c66: string, _c67:
string, _c68: string, _c69: string]
```



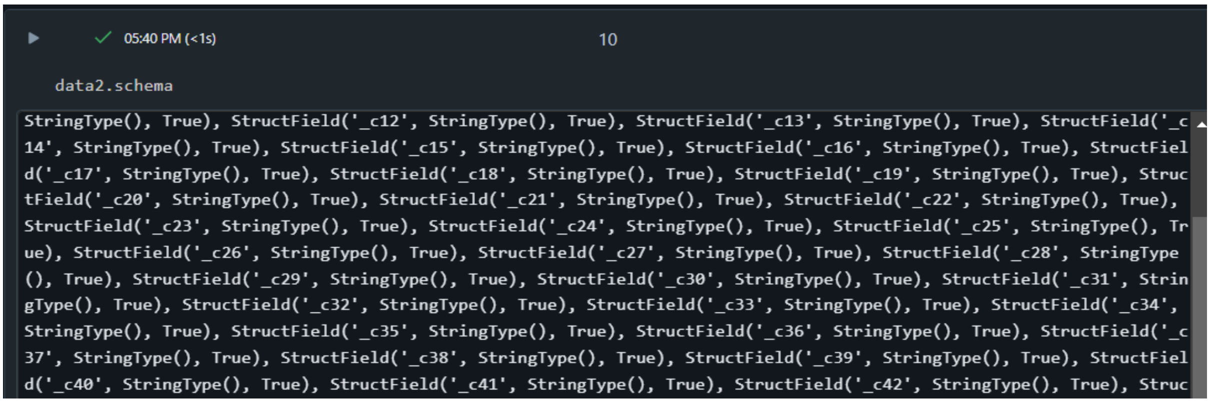
```
data2.printSchema()
```

```
 |-- _c50: string (nullable = true)
 |-- _c51: string (nullable = true)
 |-- _c52: string (nullable = true)
 |-- _c53: string (nullable = true)
 |-- _c54: string (nullable = true)
 |-- _c55: string (nullable = true)
 |-- _c56: string (nullable = true)
 |-- _c57: string (nullable = true)
 |-- _c58: string (nullable = true)
 |-- _c59: string (nullable = true)
 |-- _c60: string (nullable = true)
 |-- _c61: string (nullable = true)
 |-- _c62: string (nullable = true)
 |-- _c63: string (nullable = true)
 |-- _c64: string (nullable = true)
 |-- _c65: string (nullable = true)
 |-- _c66: string (nullable = true)
 |-- _c67: string (nullable = true)
 |-- _c68: string (nullable = true)
 |-- _c69: string (nullable = true)
```



```
data2.na.drop().show()
```

```
-------+--------------+--------------------+----------------+--------------------+--------------------+-------+-------
-----------+-----------+--------------------+-----------+--------------------+
|_c0|            _c1|                   _c2|                  _c3|_c4|  _c5| _c6|          _c7|                   _c8|
_c9|                  _c10|      _c11|              _c12|                   _c13|                   _c14|                  _c15|
_c16|                 _c17|                _c18|          _c19|                   _c20|      _c21|                  _c22|
_c23|         _c24|         _c25|          _c26|                  _c27|          _c28|  _c29|              _c30|      _c31|  _c32|
_c33|            _c34|                _c35|                 _c36|                   _c37|  _c38|                  _c39|
_c40|                 _c41|            _c42|                  _c43|                 _c44|           _c45|                  _c46
|            _c47|         _c48|                  _c49|           _c50|                    _c51|                  _c52|      _c53
|                   _c54|                 _c55|             _c56|          _c57|                    _c58|          _c59|
_c60|                 _c61|                _c62|                 _c63|    _c64|                  _c65|          _c66|
_c67|         _c68|                   _c69|
```



```
data2.schema
```

```
StringType(), True), StructField('_c12', StringType(), True), StructField('_c13', StringType(), True), StructField('_c
14', StringType(), True), StructField('_c15', StringType(), True), StructField('_c16', StringType(), True), StructFiel
d('_c17', StringType(), True), StructField('_c18', StringType(), True), StructField('_c19', StringType(), True), Struc
tField('_c20', StringType(), True), StructField('_c21', StringType(), True), StructField('_c22', StringType(), True),
StructField('_c23', StringType(), True), StructField('_c24', StringType(), True), StructField('_c25', StringType(), Tr
ue), StructField('_c26', StringType(), True), StructField('_c27', StringType(), True), StructField('_c28', StringType
(), True), StructField('_c29', StringType(), True), StructField('_c30', StringType(), True), StructField('_c31', Strin
gType(), True), StructField('_c32', StringType(), True), StructField('_c33', StringType(), True), StructField('_c34',
StringType(), True), StructField('_c35', StringType(), True), StructField('_c36', StringType(), True), StructField('_c
37', StringType(), True), StructField('_c38', StringType(), True), StructField('_c39', StringType(), True), StructFiel
d('_c40', StringType(), True), StructField('_c41', StringType(), True), StructField('_c42', StringType(), True), Struc
```