Apache PySpark by Eswar

Start Date: 05-March-2024

I have completed this project using Apache Pyspark course learnt from LinkedIn Platform. Course Link: Apache_Pyspark

Install Spark

- Google colab recently made some changes which breaks the Spark installation.
- Please use the code below where we install from the pyspark package instead

```
pip install pyspark==3.4.0
```

```
Collecting pyspark==3.4.0
  Downloading pyspark-3.4.0.tar.gz (310.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 310.8/310.8 MB 3.0 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark==3.4.0) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.4.0-py2.py3-none-any.whl size=311317122 sha256=4cc19462114507abd6dd733044d08269b4afca74a
  Stored in directory: /root/.cache/pip/wheels/7b/1b/4b/3363a1d04368e7ff0d408e57ff57966fcdf00583774e761327
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.4.0
```

```
# lets start importing sparksession and create a spark session
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
v3.4.0
Master
local[*]
AppName
pyspark-shell

Double-click (or enter) to edit

In this Project we are working on Chicago reported Crime Data

Downloading and preprocessing Chicago's Reported Crime Data

It will take sometime because we are using 1.76 Giga Bytes of Data. It is huge. Data size is increases everytime because website contains realtime which updates everyday.

```
!wget https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD
!ls -l
```

```
--2024-03-14 12:27:32--  https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD
Resolving data.cityofchicago.org (data.cityofchicago.org)... 52.206.140.199, 52.206.140.205, 52.206.68.26
Connecting to data.cityofchicago.org (data.cityofchicago.org)|52.206.140.199|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/csv]
Saving to: ‘rows.csv?accessType=DOWNLOAD’
```

```
rows.csv?accessType     [             <=>   ]   1.76G  3.03MB/s    in 9m 46s

2024-03-14 12:37:18 (3.08 MB/s) - ‘rows.csv?accessType=DOWNLOAD’ saved [1893407115]

total 1849040
-rw-r--r-- 1 root root 1893407115 Mar 14 11:02 'rows.csv?accessType=DOWNLOAD'
drwxr-xr-x 1 root root       4096 Mar 12 13:24  sample_data
```

Our File is "rows.csv", now we change that name and save as "reported-crimes.csv" file.

```
!mv rows.csv\?accessType\=DOWNLOAD reported-crimes.csv
!ls -l
```

```
total 1849040
-rw-r--r-- 1 root root 1893407115 Mar 14 11:02 reported-crimes.csv
drwxr-xr-x 1 root root       4096 Mar 12 13:24 sample_data
```

Now let's Use the show() to access the first 5 records from reported-crimes.csv

```
from pyspark.sql.functions import to_timestamp,col,lit
rc = spark.read.csv('reported-crimes.csv',header=True).withColumn('Date',to_timestamp(col('Date'),'MM/dd/yyyy hh:mm:ss a')).filter(col('Date'
rc.show(5)
```

```
+--------+-----------+-------------------+--------------------+----+-----------------+--------------------+--------------------+------+
|      ID|Case Number|               Date|               Block|IUCR|      Primary Type|         Description|Location Description|Arrest|
+--------+-----------+-------------------+--------------------+----+-----------------+--------------------+--------------------+------+
|11037294|   JA371270|2015-03-18 12:00:00|   0000X W WACKER DR|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| false|
|11645836|   JC212333|2016-05-01 00:25:00| 055XX S ROCKWELL ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| false|
|11645601|   JC212935|2014-06-01 00:01:00| 087XX S SANGAMON ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
|11646166|   JC213529|2018-09-01 00:01:00|082XX S INGLESIDE...|0810|            THEFT|           OVER $500|           RESIDENCE| false|
|11645648|   JC212959|2018-01-01 08:00:00| 024XX N MONITOR AVE|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
+--------+-----------+-------------------+--------------------+----+-----------------+--------------------+--------------------+------+
only showing top 5 rows
```

(03-03) Schemas

```
!ls
```

```
reported-crimes.csv  sample_data
```

```
rc.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- Case Number: string (nullable = true)
 |-- Date: timestamp (nullable = true)
 |-- Block: string (nullable = true)
 |-- IUCR: string (nullable = true)
 |-- Primary Type: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- Location Description: string (nullable = true)
 |-- Arrest: string (nullable = true)
 |-- Domestic: string (nullable = true)
 |-- Beat: string (nullable = true)
 |-- District: string (nullable = true)
 |-- Ward: string (nullable = true)
 |-- Community Area: string (nullable = true)
 |-- FBI Code: string (nullable = true)
 |-- X Coordinate: string (nullable = true)
 |-- Y Coordinate: string (nullable = true)
 |-- Year: string (nullable = true)
 |-- Updated On: string (nullable = true)
 |-- Latitude: string (nullable = true)
 |-- Longitude: string (nullable = true)
 |-- Location: string (nullable = true)
```

We import the types from pyspark and note down all the columns and

```
from pyspark.sql.types import StructType,StructField, StringType, TimestampType, BooleanType, DoubleType, IntegerType
```

```
rc.columns
```

```
['ID',
 'Case Number',
 'Date',
 'Block',
 'IUCR',
 'Primary Type',
 'Description',
 'Location Description',
 'Arrest',
 'Domestic',
 'Beat',
 'District',
 'Ward',
 'Community Area',
 'FBI Code',
 'X Coordinate',
 'Y Coordinate',
 'Year',
 'Updated On',
 'Latitude',
 'Longitude',
 'Location']
```

```
schema = StructType([
    StructField('ID', StringType(), True),
    StructField('Case Number', StringType(), True),
    StructField('Date', TimestampType(), True),
    StructField('Block', StringType(), True),
    StructField('IUCR', StringType(), True),
    StructField('Primary Type', StringType(), True),
    StructField('Location Description', StringType(), True),
    StructField('Arrest', StringType(), True),
    StructField('Domestic', BooleanType(), True),
    StructField('District', StringType(), True),
    StructField('Ward', StringType(), True),
    StructField('Community Area', StringType(), True),
    StructField('FBI Code', StringType(), True),
    StructField('X Coordinate', StringType(), True),
    StructField('Y Coordinate', StringType(), True),
    StructField('Year', IntegerType(), True),
    StructField('Updated On', StringType(), True),
    StructField('Latitude', DoubleType(), True),
    StructField('Longitude', DoubleType(), True),
    StructField('Location', StringType(), True)
])
```

Lets add schema to reported-crimes.csv file and check the schema.

```
rc = spark.read.csv('reported-crimes.csv',schema=schema)
rc.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- Case Number: string (nullable = true)
 |-- Date: timestamp (nullable = true)
 |-- Block: string (nullable = true)
 |-- IUCR: string (nullable = true)
 |-- Primary Type: string (nullable = true)
 |-- Location Description: string (nullable = true)
 |-- Arrest: string (nullable = true)
 |-- Domestic: boolean (nullable = true)
 |-- District: string (nullable = true)
 |-- Ward: string (nullable = true)
 |-- Community Area: string (nullable = true)
 |-- FBI Code: string (nullable = true)
 |-- X Coordinate: string (nullable = true)
 |-- Y Coordinate: string (nullable = true)
```

```
 |-- Year: integer (nullable = true)
 |-- Updated On: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Location: string (nullable = true)
```

Here we see few columns as null because some datatypes are not adjusted to fields we have given.

```
rc.show(5)
```

```
+--------+-----------+----+--------------------+----+-----------------+--------------------+--------------------+--------+--------+----
|      ID|Case Number|Date|               Block|IUCR|      Primary Type|Location Description|              Arrest|Domestic|District|Ward
+--------+-----------+----+--------------------+----+-----------------+--------------------+--------------------+--------+--------+----
|      ID|Case Number|null|               Block|IUCR|      Primary Type|         Description|Location Description|    null|Domestic|Beat
|11037294|   JA371270|null|   0000X W WACKER DR|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK|   false|   false|0111
|11646293|   JC213749|null|023XX N LOCKWOOD AVE|1154|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT|   false|   false|2515
|11645836|   JC212333|null| 055XX S ROCKWELL ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null|   false|   false|0824
|11645959|   JC211511|null|  045XX N ALBANY AVE|2820|     OTHER OFFENSE|    TELEPHONE THREAT|           RESIDENCE|   false|   false|1724
+--------+-----------+----+--------------------+----+-----------------+--------------------+--------------------+--------+--------+----
only showing top 5 rows
```

(03-04) Working with columns

Display only the first 5 rows of the column name IUCR

```
rc.select('IUCR').show(5)
```

```
+----+
|IUCR|
+----+
|IUCR|
|1153|
|1154|
|1153|
|2820|
+----+
only showing top 5 rows
```

```
#Other way to display 5 records
rc.select(col('IUCR')).show(5)
```

```
+----+
|IUCR|
+----+
|IUCR|
|1153|
|1154|
|1153|
|2820|
+----+
only showing top 5 rows
```

Display only the first 4 rows of the column names Case Number, Date and Arrest

```
rc.select('Case Number','Date','Arrest').show(4)
```

```
+-----------+----+--------------------+
|Case Number|Date|              Arrest|
+-----------+----+--------------------+
|Case Number|null|Location Description|
|   JA371270|null|                BANK|
|   JC213749|null|           APARTMENT|
|   JC212333|null|                null|
+-----------+----+--------------------+
```

```
only showing top 4 rows
```

Add a column with name One, with entries all 1s

#In PySpark, the lit() function is used to create a column expression with a literal value. It is commonly used when you want to add a constant value as a new column or as part of a transformation on an existing column within a DataFrame.

```
# Here we import function lit
from pyspark.sql.functions import lit
rc.withColumn('One',lit(1)).show(5)
```

```
+--------+-----------+----+--------------------+----+-----------------+--------------------+--------------------+--------+--------+----
|      ID|Case Number|Date|               Block|IUCR|      Primary Type|Location Description|              Arrest|Domestic|District|Ward
+--------+-----------+----+--------------------+----+-----------------+--------------------+--------------------+--------+--------+----
|      ID|Case Number|null|               Block|IUCR|      Primary Type|         Description|Location Description|    null|Domestic|Beat
|11037294|   JA371270|null|   0000X W WACKER DR|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK|   false|   false|0111
|11646293|   JC213749|null|023XX N LOCKWOOD AVE|1154|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT|   false|   false|2515
|11645836|   JC212333|null| 055XX S ROCKWELL ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null|   false|   false|0824
|11645959|   JC211511|null|  045XX N ALBANY AVE|2820|     OTHER OFFENSE|    TELEPHONE THREAT|           RESIDENCE|   false|   false|1724
+--------+-----------+----+--------------------+----+-----------------+--------------------+--------------------+--------+--------+----
only showing top 5 rows
```

Remove the column IUCR

```
rc=rc.drop('IUCR')
```

```
rc.show(5)
```

```
+--------+-----------+----+--------------------+-----------------+--------------------+--------------------+--------+--------+----+----
|      ID|Case Number|Date|               Block|      Primary Type|Location Description|              Arrest|Domestic|District|Ward|Comm
+--------+-----------+----+--------------------+-----------------+--------------------+--------------------+--------+--------+----+----
|      ID|Case Number|null|               Block|      Primary Type|         Description|Location Description|    null|Domestic|Beat|
|11037294|   JA371270|null|   0000X W WACKER DR|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK|   false|   false|0111|
|11646293|   JC213749|null|023XX N LOCKWOOD AVE|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT|   false|   false|2515|
|11645836|   JC212333|null| 055XX S ROCKWELL ST|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null|   false|   false|0824|
|11645959|   JC211511|null|  045XX N ALBANY AVE|     OTHER OFFENSE|    TELEPHONE THREAT|           RESIDENCE|   false|   false|1724|
+--------+-----------+----+--------------------+-----------------+--------------------+--------------------+--------+--------+----+----
only showing top 5 rows
```

(03-05) Working with rows

```
rc.show(5)
```

```
+--------+-----------+----+--------------------+-----------------+--------------------+--------------------+--------+--------+----+----
|      ID|Case Number|Date|               Block|      Primary Type|Location Description|              Arrest|Domestic|District|Ward|Comm
+--------+-----------+----+--------------------+-----------------+--------------------+--------------------+--------+--------+----+----
|      ID|Case Number|null|               Block|      Primary Type|         Description|Location Description|    null|Domestic|Beat|
|11037294|   JA371270|null|   0000X W WACKER DR|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK|   false|   false|0111|
|11646293|   JC213749|null|023XX N LOCKWOOD AVE|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT|   false|   false|2515|
|11645836|   JC212333|null| 055XX S ROCKWELL ST|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null|   false|   false|0824|
|11645959|   JC211511|null|  045XX N ALBANY AVE|     OTHER OFFENSE|    TELEPHONE THREAT|           RESIDENCE|   false|   false|1724|
+--------+-----------+----+--------------------+-----------------+--------------------+--------------------+--------+--------+----+----
only showing top 5 rows
```

Add the reported crimes for an additional day, 12-Nov-2018, to our dataset.

```
one_day = spark.read.csv('reported-crimes.csv',header=True).withColumn('Date',to_timestamp(col('Date'),'MM/dd/yyyy hh:mm:ss a')).filter(col(
one_day.count()
```

```
4
```

```
one_day.show(4)
```

```
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+-
|      ID|Case Number|               Date|              Block|IUCR|      Primary Type|         Description|Location Description|Arrest|D
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+-
|13358766|   JH140578|2018-11-12 00:00:00|    008XX E 63RD ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|          APARTMENT| false|
|11540042|   JB559262|2018-11-12 00:00:00|010XX N DEARBORN ST|1140|DECEPTIVE PRACTICE|        EMBEZZLEMENT|  CONVENIENCE STORE|  true|
|11516594|   JB528186|2018-11-12 00:00:00|049XX S PRAIRIE AVE|2826|     OTHER OFFENSE|HARASSMENT BY ELE...|               OTHER| false|
|11505149|   JB513151|2018-11-12 00:00:00| 003XX S WHIPPLE ST|0810|            THEFT|           OVER $500|              STREET| false|
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+-
```

```
#Here we need to create rc again because we have change rc in above operation,
#so union will return error due to mismatch coulumns.

from pyspark.sql.functions import to_timestamp,col,lit
rc = spark.read.csv('reported-crimes.csv',header=True).withColumn('Date',to_timestamp(col('Date'),'MM/dd/yyyy hh:mm:ss a')).filter(col('Date
rc.show(5)
```

```
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+
|      ID|Case Number|               Date|              Block|IUCR|      Primary Type|         Description|Location Description|Arrest|
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+
|11037294|   JA371270|2015-03-18 12:00:00|   0000X W WACKER DR|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| false|
|11645836|   JC212333|2016-05-01 00:25:00| 055XX S ROCKWELL ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| false|
|11645601|   JC212935|2014-06-01 00:01:00| 087XX S SANGAMON ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
|11646166|   JC213529|2018-09-01 00:01:00|082XX S INGLESIDE...|0810|            THEFT|           OVER $500|           RESIDENCE| false|
|11645648|   JC212959|2018-01-01 08:00:00| 024XX N MONITOR AVE|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+
only showing top 5 rows
```

```
#Now lets add these data in bottom of dataframe
rc.union(one_day).show(5)
```

```
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+
|      ID|Case Number|               Date|              Block|IUCR|      Primary Type|         Description|Location Description|Arrest|
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+
|11037294|   JA371270|2015-03-18 12:00:00|   0000X W WACKER DR|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| false|
|11645836|   JC212333|2016-05-01 00:25:00| 055XX S ROCKWELL ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| false|
|11645601|   JC212935|2014-06-01 00:01:00| 087XX S SANGAMON ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
|11646166|   JC213529|2018-09-01 00:01:00|082XX S INGLESIDE...|0810|            THEFT|           OVER $500|           RESIDENCE| false|
|11645648|   JC212959|2018-01-01 08:00:00| 024XX N MONITOR AVE|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
+--------+-----------+-------------------+-------------------+----+-----------------+--------------------+--------------------+------+
only showing top 5 rows
```

```
#Union is completed but we can't see excepted data, because our data is present
#in bottom of the dataframe
#Let's sort data in ascending oder according to Date
rc.orderBy('Date',ascending=False).show(5)
```

```
+--------+-----------+-------------------+-------------------+----+--------------------+--------------------+--------------------+------
|      ID|Case Number|               Date|              Block|IUCR|        Primary Type|         Description|Location Description|Arrest
+--------+-----------+-------------------+-------------------+----+--------------------+--------------------+--------------------+------
|11513303|   JB523990|2018-11-11 00:00:00| 007XX S CICERO AVE|0281|CRIMINAL SEXUAL A...|      NON-AGGRAVATED|              STREET| false
|11507744|   JB516509|2018-11-11 00:00:00|  033XX S DAMEN AVE|0820|               THEFT|      $500 AND UNDER|           RESIDENCE| false
|11505877|   JB514053|2018-11-11 00:00:00|040XX S PRAIRIE AVE|0820|               THEFT|      $500 AND UNDER|               OTHER| false
|11504529|   JB512336|2018-11-11 00:00:00|    026XX W 21ST ST|1320|     CRIMINAL DAMAGE|          TO VEHICLE|              STREET| false
|11505581|   JB513620|2018-11-11 00:00:00| 049XX S KEELER AVE|0486|             BATTERY|DOMESTIC BATTERY ...|           RESIDENCE| false
+--------+-----------+-------------------+-------------------+----+--------------------+--------------------+--------------------+------
only showing top 5 rows
```

```
#lets get our result
rc.union(one_day).orderBy('Date',ascending=False).show(5)
```

```
+--------+-----------+-------------------+-------------------+----+--------------------+--------------------+--------------------+------+
|      ID|Case Number|               Date|              Block|IUCR|        Primary Type|         Description|Location Description|Arrest|
+--------+-----------+-------------------+-------------------+----+--------------------+--------------------+--------------------+------+
|13358766|   JH140578|2018-11-12 00:00:00|    008XX E 63RD ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT| false|
|11540042|   JB559262|2018-11-12 00:00:00|010XX N DEARBORN ST|1140|  DECEPTIVE PRACTICE|        EMBEZZLEMENT|   CONVENIENCE STORE|  true|
|11516594|   JB528186|2018-11-12 00:00:00|049XX S PRAIRIE AVE|2826|       OTHER OFFENSE|HARASSMENT BY ELE...|               OTHER| false|
|11505149|   JB513151|2018-11-12 00:00:00| 003XX S WHIPPLE ST|0810|               THEFT|           OVER $500|              STREET| false|
|11513303|   JB523990|2018-11-11 00:00:00| 007XX S CICERO AVE|0281|CRIMINAL SEXUAL A...|      NON-AGGRAVATED|              STREET| false|
+--------+-----------+-------------------+-------------------+----+--------------------+--------------------+--------------------+------+
only showing top 5 rows
```

What are the top 10 number of reported crimes by Primary type, in descending order of occurence?

```
#Here we check the count of each primary type
rc.groupBy('Primary Type').count().show()
```

```
+--------------------+-------+
|        Primary Type|  count|
+--------------------+-------+
|OFFENSE INVOLVING...|  46891|
|CRIMINAL SEXUAL A...|   1422|
|            STALKING|   3388|
|PUBLIC PEACE VIOL...|  47785|
|           OBSCENITY|    586|
|               ARSON|  11157|
|            GAMBLING|  14422|
|   CRIMINAL TRESPASS| 193372|
|             ASSAULT| 418522|
|LIQUOR LAW VIOLATION|  14068|
| MOTOR VEHICLE THEFT| 314136|
|               THEFT|1418530|
|             BATTERY|1232293|
|             ROBBERY| 255604|
|            HOMICIDE|   9476|
|    PUBLIC INDECENCY|    161|
| CRIM SEXUAL ASSAULT|  26373|
|   HUMAN TRAFFICKING|     48|
|        INTIMIDATION|   3937|
|        PROSTITUTION|  68327|
+--------------------+-------+
only showing top 20 rows
```

```
#Now we fetch data in desceding order with top 10.
rc.groupBy('Primary Type').count().orderBy('count',ascending=False).show(10)
```

```
+-------------------+-------+
|       Primary Type|  count|
+-------------------+-------+
|              THEFT|1418530|
|            BATTERY|1232293|
|    CRIMINAL DAMAGE| 771523|
|          NARCOTICS| 711778|
|      OTHER OFFENSE| 419048|
|            ASSAULT| 418522|
|           BURGLARY| 388042|
|MOTOR VEHICLE THEFT| 314136|
| DECEPTIVE PRACTICE| 267312|
|            ROBBERY| 255604|
+-------------------+-------+
only showing top 10 rows
```

(03-06) Challenge

What percentage of reported crimes resulted in an arrest?

```
a=rc.filter(col('Arrest')=='true').count() # Caliculated total number of arrested people
b=rc.select(col('Arrest')).count()  # Caliculated total number of people
result=(a/b)*100 # Caliculated percentage of reported crimes resulted in a arrest
print("Total Number of arrested people = ",a)
print("Total number of people =",b)
print("Percentage of reported crimes resulted in an arrest =",result)
```

```
Total Number of arrested people =  1875314
Total number of people = 6756974
Percentage of reported crimes resulted in an arrest = 27.7537548612737
```

What are the top 3 locations for reported crimes?

```
rc.groupBy('Location Description').count().orderBy('count',ascending=False).show(3)
```

```
+--------------------+-------+
|Location Description|  count|
+--------------------+-------+
|              STREET|1770639|
|           RESIDENCE|1146377|
|           APARTMENT| 699284|
+--------------------+-------+
only showing top 3 rows
```

(04-01) Built-in functions

```
from pyspark.sql import functions
```

```
print(dir(functions))
```

```
['Any', 'ArrayType', 'Callable', 'Column', 'DataFrame', 'DataType', 'Dict', 'Iterable', 'JVMView', 'List', 'Optional', 'PandasUDFType',
```

String functions

Display the Primary Type column in lower and upper characters, and the first 4 characters of the column

```
rc.show()
```

```
+--------+-----------+------------------+--------------------+----+--------------------+--------------------+--------------------+-----
|      ID|Case Number|              Date|               Block|IUCR|        Primary Type|         Description|Location Description|Arres
+--------+-----------+------------------+--------------------+----+--------------------+--------------------+--------------------+-----
|11037294|   JA371270|2015-03-18 12:00:00|   0000X W WACKER DR|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| fals
|11645836|   JC212333|2016-05-01 00:25:00| 055XX S ROCKWELL ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
|11645601|   JC212935|2014-06-01 00:01:00| 087XX S SANGAMON ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11646166|   JC213529|2018-09-01 00:01:00|082XX S INGLESIDE...|0810|               THEFT|           OVER $500|           RESIDENCE| fals
|11645648|   JC212959|2018-01-01 08:00:00| 024XX N MONITOR AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11645557|   JC212685|2018-04-01 00:01:00|  080XX S VERNON AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11645527|   JC212744|2015-02-02 10:00:00|  069XX W ARCHER AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|               OTHER| fals
|11645833|   JC213044|2012-05-05 12:25:00|     057XX W OHIO ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
| 4144897|   HL474854|2005-07-10 15:00:00| 062XX S ABERDEEN ST|0430|             BATTERY|AGGRAVATED: OTHER...|              STREET| fals
| 1744168|    G553545|2001-09-15 02:00:00|     013XX W POLK ST|0460|             BATTERY|              SIMPLE|              STREET| fals
|11615821|   JC176668|2016-01-01 12:00:00| 054XX N NATCHEZ AVE|1195|  DECEPTIVE PRACTICE|FINAN EXPLOIT-ELD...|           RESIDENCE| fals
|11641400|   JC207897|2018-07-26 13:00:00|  041XX N KEELER AVE|1130|  DECEPTIVE PRACTICE|FRAUD OR CONFIDEN...|           RESIDENCE| fals
|11646766|   JC214025|2018-01-01 09:10:00| 019XX N KENMORE AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
|11646447|   JC213946|2008-10-24 14:30:00|036XX N NARRAGANS...|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT| fals
|11646550|   JC214099|2018-10-04 12:00:00|   092XX S PERRY AVE|1120|  DECEPTIVE PRACTICE|             FORGERY|               OTHER| fals
| 4229528|   HL545852|2005-08-12 23:00:00|063XX S COTTAGE G...|3730|INTERFERENCE WITH...| OBSTRUCTING JUSTICE|            SIDEWALK|  tru
|11647549|   JC214237|2018-06-07 15:00:00|102XX S EGGLESTON...|0820|               THEFT|      $500 AND UNDER|               OTHER| fals
|11031104|   JA362043|2008-07-24 00:01:00| 031XX W FILLMORE ST|1563|         SEX OFFENSE|CRIMINAL SEXUAL A...|           APARTMENT| fals
|11648173|   JC216048|2018-11-01 00:00:00|062XX S MARSHFIEL...|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT| fals
|11445072|   JB415614|2018-08-30 19:45:00| 031XX W HARRISON ST|2027|           NARCOTICS|         POSS: CRACK|POLICE FACILITY/V...|  tru
+--------+-----------+------------------+--------------------+----+--------------------+--------------------+--------------------+-----
only showing top 20 rows
```

```
from pyspark.sql.functions import upper,lower,substring
```

```
rc.select(upper(col('Primary Type')),lower(col('Primary Type')),substring(col('Primary Type'),1,5)).show(5)
```

```
+------------------+------------------+---------------------------+
|upper(Primary Type)|lower(Primary Type)|substring(Primary Type, 1, 5)|
+------------------+------------------+---------------------------+
| DECEPTIVE PRACTICE| deceptive practice|                      DECEP|
| DECEPTIVE PRACTICE| deceptive practice|                      DECEP|
| DECEPTIVE PRACTICE| deceptive practice|                      DECEP|
|              THEFT|              theft|                      THEFT|
| DECEPTIVE PRACTICE| deceptive practice|                      DECEP|
+------------------+------------------+---------------------------+
only showing top 5 rows
```

Numeric functions

Show the oldest date and the most recent date

```
from pyspark.sql.functions import min,max
```

```
rc.select(max(col('Date')),min(col('Date'))).show(1)
```

```
+-------------------+-------------------+
|          max(Date)|          min(Date)|
+-------------------+-------------------+
|2018-11-11 00:00:00|2001-01-01 00:00:00|
+-------------------+-------------------+
```

Date

What is 3 days earlier that the oldest date and 3 days later than the most recent date?

```
from pyspark.sql.functions import date_add,date_sub
```

```
rc.select(date_sub(min(col('Date')),3),date_add(max(col('Date')),3)).show()
```

```
+----------------------+----------------------+
|date_sub(min(Date), 3)|date_add(max(Date), 3)|
+----------------------+----------------------+
|            2000-12-29|            2018-11-14|
+----------------------+----------------------+
```

(04-02) Working with dates

```
from pyspark.sql.functions import to_date, to_timestamp, lit
```

2019-12-25 13:30:00

```
df = spark.createDataFrame([('2019-12-25 13:30:00',)],['Christmas'])
df.show(1)
```

```
+-------------------+
|          Christmas|
+-------------------+
|2019-12-25 13:30:00|
+-------------------+
```

```
df.select(to_date(col('Christmas'),'yyyy-MM-dd HH:mm:ss'), to_timestamp(col('Christmas'),'yyyy-MM-dd HH:mm:ss')).show(1)
```

```
+----------------------------------------+-------------------------------------------+
|to_date(Christmas, yyyy-MM-dd HH:mm:ss)|to_timestamp(Christmas, yyyy-MM-dd HH:mm:ss)|
+----------------------------------------+-------------------------------------------+
|                              2019-12-25|                        2019-12-25 13:30:00|
```

```
+-------------------------------------+-----------------------------------------+
```

25/Dec/2019 13:30:00

```
df = spark.createDataFrame([('25/Dec/2019 13:30:00',)],['Christmas'])
df.show(1)
```

```
+--------------------+
|           Christmas|
+--------------------+
|25/Dec/2019 13:30:00|
+--------------------+
```

```
df.select(to_date(col('Christmas'),'dd/MMM/yyyy HH:mm:ss'),to_timestamp(col('Christmas'),'dd/MMM/yyyy HH:mm:ss')).show(1)
```

```
+----------------------------------------+-------------------------------------------+
|to_date(Christmas, dd/MMM/yyyy HH:mm:ss)|to_timestamp(Christmas, dd/MMM/yyyy HH:mm:ss)|
+----------------------------------------+-------------------------------------------+
|                              2019-12-25|                        2019-12-25 13:30:00|
+----------------------------------------+-------------------------------------------+
```

12/25/2019 01:30:00 PM

```
df = spark.createDataFrame([('12/25/2019 01:30:00 PM',)],['Christmas'])
df.show(1,truncate=False)
```

```
+----------------------+
|Christmas             |
+----------------------+
|12/25/2019 01:30:00 PM|
+----------------------+
```

```
nrc=spark.read.csv('reported-crimes.csv',header=True)
nrc.show(5,truncate=False)
```

```
+--------+-----------+----------------------+--------------------+----+-----------------+-------------------------------------+------
|ID      |Case Number|Date                  |Block               |IUCR|Primary Type     |Description                          |Locati
+--------+-----------+----------------------+--------------------+----+-----------------+-------------------------------------+------
|11037294|JA371270   |03/18/2015 12:00:00 PM|0000X W WACKER DR   |1153|DECEPTIVE PRACTICE|FINANCIAL IDENTITY THEFT OVER $ 300    |BANK
|11646293|JC213749   |12/20/2018 03:00:00 PM|023XX N LOCKWOOD AVE|1154|DECEPTIVE PRACTICE|FINANCIAL IDENTITY THEFT $300 AND UNDER|APARTM
|11645836|JC212333   |05/01/2016 12:25:00 AM|055XX S ROCKWELL ST |1153|DECEPTIVE PRACTICE|FINANCIAL IDENTITY THEFT OVER $ 300    |null
|11645959|JC211511   |12/20/2018 04:00:00 PM|045XX N ALBANY AVE  |2820|OTHER OFFENSE     |TELEPHONE THREAT                     |RESIDE
|11645601|JC212935   |06/01/2014 12:01:00 AM|087XX S SANGAMON ST |1153|DECEPTIVE PRACTICE|FINANCIAL IDENTITY THEFT OVER $ 300    |RESIDE
+--------+-----------+----------------------+--------------------+----+-----------------+-------------------------------------+------
only showing top 5 rows
```

(04-03) Joins

Download police station data

```
rc.show(4)
```

```
+--------+-----------+-------------------+--------------------+----+-----------------+--------------------+--------------------+------+
|      ID|Case Number|               Date|               Block|IUCR|      Primary Type|         Description|Location Description|Arrest|
+--------+-----------+-------------------+--------------------+----+-----------------+--------------------+--------------------+------+
|11037294|   JA371270|2015-03-18 12:00:00|   0000X W WACKER DR|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| false|
|11645836|   JC212333|2016-05-01 00:25:00| 055XX S ROCKWELL ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| false|
|11645601|   JC212935|2014-06-01 00:01:00| 087XX S SANGAMON ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
|11646166|   JC213529|2018-09-01 00:01:00|082XX S INGLESIDE...|0810|             THEFT|           OVER $500|           RESIDENCE| false|
+--------+-----------+-------------------+--------------------+----+-----------------+--------------------+--------------------+------+
only showing top 4 rows
```

Double-click (or enter) to edit

```
#We have downloaded the police-station data from city og chicago website


ps=spark.read.csv('/content/police-stations.csv',header=True)
ps.show(5)
```

```
+------------+-------------+--------------------+-------+-----+-----+--------------------+------------+------------+------------+--------
|    DISTRICT|DISTRICT NAME|             ADDRESS|   CITY|STATE|  ZIP|             WEBSITE|       PHONE|         FAX|         TTY|X COORDI
+------------+-------------+--------------------+-------+-----+-----+--------------------+------------+------------+------------+--------
|Headquarters| Headquarters|3510 S Michigan Ave|Chicago|   IL|60653|http://home.chica...|        null|        null|        null| 1177731
|          18|   Near North| 1160 N Larrabee St|Chicago|   IL|60610|http://home.chica...|312-742-5870|312-742-5771|312-742-5773| 1172086
|          19|    Town Hall|   850 W Addison St|Chicago|   IL|60613|http://home.chica...|312-744-8320|312-744-4481|312-744-8011| 1169736
|          20|      Lincoln| 5400 N Lincoln Ave|Chicago|   IL|60625|http://home.chica...|312-742-8714|312-742-8803|312-742-8841| 1158399
|          22|  Morgan Park|1900 W Monterey Ave|Chicago|   IL|60643|http://home.chica...|312-745-0710|312-745-0814|312-745-0569| 1165825
+------------+-------------+--------------------+-------+-----+-----+--------------------+------------+------------+------------+--------
only showing top 5 rows
```

The reported crimes dataset has only the district number. Add the district name by joining with the police station dataset

```
rc.cache()
rc.count()
```

```
ERROR:root:KeyboardInterrupt while sending command.
Traceback (most recent call last):
  File "/usr/local/lib/python3.10/dist-packages/py4j/java_gateway.py", line 1038, in send_command
    response = connection.send_command(command)
  File "/usr/local/lib/python3.10/dist-packages/py4j/clientserver.py", line 511, in send_command
    answer = smart_decode(self.stream.readline()[:-1])
  File "/usr/lib/python3.10/socket.py", line 705, in readinto
    return self._sock.recv_into(b)
KeyboardInterrupt
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-107-d50cbf834367> in <cell line: 2>()
      1 rc.cache()
----> 2 rc.count()

 4 frames
/usr/lib/python3.10/socket.py in readinto(self, b)
    703         while True:
    704             try:
--> 705                 return self._sock.recv_into(b)
    706             except timeout:
    707                 self._timeout_occurred = True

KeyboardInterrupt:
```

```
ps.select(col('DISTRICT')).distinct().show(30)
```

```
+--------------------+
|            DISTRICT|
+--------------------+
|                   7|
|                  15|
|                  11|
|                   3|
|                   8|
|                  22|
|                  16|
|                   5|
|                  18|
|                  17|
|                   6|
|                  19|
|                  25|
|        Headquarters|
|                  24|
|                   9|
|                   1|
|                  20|
|                  10|
|                   4|
|                  12|
```

```
|                 14|
|                  2|
|",Chicago,IL,6060...|
+--------------------+
```

```
rc.select('District').distinct().show(30)
```

```
+--------+
|District|
+--------+
|     009|
|     012|
|     024|
|    null|
|     031|
|     015|
|     006|
|     019|
|     020|
|     011|
|     025|
|     003|
|     005|
|     016|
|     018|
|     008|
|     022|
|     001|
|     014|
|     010|
|     004|
|     017|
|     007|
|     002|
|     021|
+--------+
```

```
#We do padding because we have leading zeros in one dataset.
from pyspark.sql.functions import lpad
```

```
ps.select(lpad(col('DISTRICT'),3,'0')).show()
```

```
+--------------------+
|lpad(DISTRICT, 3, 0)|
+--------------------+
|                 Hea|
|                 018|
|                 019|
|                 020|
|                 022|
|                 024|
|                 025|
|                 001|
|                 002|
|                 003|
|                 004|
|                 005|
|                 006|
|                 007|
|                 008|
|                 009|
|                 010|
|                 011|
|                 012|
|                 ",C|
+--------------------+
only showing top 20 rows
```

```
ps = ps.withColumn('Format_district',lpad(col('DISTRICT'),3,'0'))
ps.show(5)
```

```
-----+-----+--------------------+------------+------------+------------+------------+------------+-----------+------------+--------------
STATE|  ZIP|             WEBSITE|       PHONE|         FAX|         TTY|X COORDINATE|Y COORDINATE|   LATITUDE|   LONGITUDE|            LC
-----+-----+--------------------+------------+------------+------------+------------+------------+-----------+------------+--------------
   IL|60653|http://home.chica...|        null|        null|        null| 1177731.401| 1881697.404|41.83070169|-87.62339535|(41.830701687
   IL|60610|http://home.chica...|312-742-5870|312-742-5771|312-742-5773| 1172080.029| 1908086.527|41.90324165|-87.64335214|(41.903241653
   IL|60613|http://home.chica...|312-744-8320|312-744-4481|312-744-8011| 1169730.744| 1924160.317|41.94740046|-87.65151202|(41.947400456
```

```
IL|60625|http://home.chica...|312-742-8714|312-742-8803|312-742-8841| 1158399.146| 1935788.826|41.97954951|-87.69284451|(41.9795495131
IL|60643|http://home.chica...|312-745-0710|312-745-0814|312-745-0569| 1165825.476| 1830851.333|41.69143478|-87.66852039|(41.6914347795
-----+-----+--------------------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-------------
```

```
rc.join(ps, rc.District == ps.Format_district, 'left_outer').show()
```

```
+--------+-----------+-------------------+--------------------+----+--------------------+--------------------+--------------------+-----
|      ID|Case Number|               Date|               Block|IUCR|        Primary Type|         Description|Location Description|Arres
+--------+-----------+-------------------+--------------------+----+--------------------+--------------------+--------------------+-----
|11037294|   JA371270|2015-03-18 12:00:00|   0000X W WACKER DR|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| fals
|11645836|   JC212333|2016-05-01 00:25:00| 055XX S ROCKWELL ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
|11645601|   JC212935|2014-06-01 00:01:00| 087XX S SANGAMON ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11646166|   JC213529|2018-09-01 00:01:00|082XX S INGLESIDE...|0810|               THEFT|           OVER $500|           RESIDENCE| fals
|11645648|   JC212959|2018-01-01 08:00:00| 024XX N MONITOR AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11645557|   JC212685|2018-04-01 00:01:00|  080XX S VERNON AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11645527|   JC212744|2015-02-02 10:00:00|  069XX W ARCHER AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|               OTHER| fals
|11645833|   JC213044|2012-05-05 12:25:00|     057XX W OHIO ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
| 4144897|   HL474854|2005-07-10 15:00:00| 062XX S ABERDEEN ST|0430|             BATTERY|AGGRAVATED: OTHER...|              STREET| fals
| 1744168|    G553545|2001-09-15 02:00:00|     013XX W POLK ST|0460|             BATTERY|              SIMPLE|              STREET| fals
|11615821|   JC176668|2016-01-01 12:00:00| 054XX N NATCHEZ AVE|1195|  DECEPTIVE PRACTICE|FINAN EXPLOIT-ELD...|           RESIDENCE| fals
|11641400|   JC207897|2018-07-26 13:00:00|  041XX N KEELER AVE|1130|  DECEPTIVE PRACTICE|FRAUD OR CONFIDEN...|           RESIDENCE| fals
|11646766|   JC214025|2018-01-01 09:10:00| 019XX N KENMORE AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
|11646447|   JC213946|2008-10-24 14:30:00|036XX N NARRAGANS...|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT| fals
|11646550|   JC214099|2018-10-04 12:00:00|   092XX S PERRY AVE|1120|  DECEPTIVE PRACTICE|             FORGERY|               OTHER| fals
| 4229528|   HL545852|2005-08-12 23:00:00|063XX S COTTAGE G...|3730|INTERFERENCE WITH...| OBSTRUCTING JUSTICE|            SIDEWALK|  tru
|11647549|   JC214237|2018-06-07 15:00:00|102XX S EGGLESTON...|0820|               THEFT|      $500 AND UNDER|               OTHER| fals
|11031104|   JA362043|2008-07-24 00:01:00| 031XX W FILLMORE ST|1563|         SEX OFFENSE|CRIMINAL SEXUAL A...|           APARTMENT| fals
|11648173|   JC216048|2018-11-01 00:00:00|062XX S MARSHFIEL...|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT| fals
|11445072|   JB415614|2018-08-30 19:45:00| 031XX W HARRISON ST|2027|           NARCOTICS|         POSS: CRACK|POLICE FACILITY/V...|  tru
+--------+-----------+-------------------+--------------------+----+--------------------+--------------------+--------------------+-----
only showing top 20 rows
```

```
ps.columns
```

```
['DISTRICT',
 'DISTRICT NAME',
 'ADDRESS',
 'CITY',
 'STATE',
 'ZIP',
 'WEBSITE',
 'PHONE',
 'FAX',
 'TTY',
 'X COORDINATE',
 'Y COORDINATE',
 'LATITUDE',
 'LONGITUDE',
 'LOCATION',
 'Format_district']
```

```
rc.join(ps,rc.District == ps.Format_district, 'left_outer').drop(
 'ADDRESS',
 'CITY',
 'STATE',
 'ZIP',
 'WEBSITE',
 'PHONE',
 'FAX',
 'TTY',
 'X COORDINATE',
 'Y COORDINATE',
 'LATITUDE',
 'LONGITUDE',
 'LOCATION'
 ).show()
```

```
+--------+-----------+-------------------+--------------------+----+--------------------+--------------------+--------------------+-----
|      ID|Case Number|               Date|               Block|IUCR|        Primary Type|         Description|Location Description|Arres
+--------+-----------+-------------------+--------------------+----+--------------------+--------------------+--------------------+-----
|11037294|   JA371270|2015-03-18 12:00:00|   0000X W WACKER DR|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| fals
|11645836|   JC212333|2016-05-01 00:25:00| 055XX S ROCKWELL ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
|11645601|   JC212935|2014-06-01 00:01:00| 087XX S SANGAMON ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11646166|   JC213529|2018-09-01 00:01:00|082XX S INGLESIDE...|0810|               THEFT|           OVER $500|           RESIDENCE| fals
```

```
|11645648|   JC212959|2018-01-01 08:00:00| 024XX N MONITOR AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11645557|   JC212685|2018-04-01 00:01:00|  080XX S VERNON AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| fals
|11645527|   JC212744|2015-02-02 10:00:00|  069XX W ARCHER AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|               OTHER| fals
|11645833|   JC213044|2012-05-05 12:25:00|     057XX W OHIO ST|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
| 4144897|   HL474854|2005-07-10 15:00:00| 062XX S ABERDEEN ST|0430|              BATTERY|AGGRAVATED: OTHER...|              STREET| fals
| 1744168|    G553545|2001-09-15 02:00:00|     013XX W POLK ST|0460|              BATTERY|              SIMPLE|              STREET| fals
|11615821|   JC176668|2016-01-01 12:00:00| 054XX N NATCHEZ AVE|1195|  DECEPTIVE PRACTICE|FINAN EXPLOIT-ELD...|           RESIDENCE| fals
|11641400|   JC207897|2018-07-26 13:00:00|  041XX N KEELER AVE|1130|  DECEPTIVE PRACTICE|FRAUD OR CONFIDEN...|           RESIDENCE| fals
|11646766|   JC214025|2018-01-01 09:10:00| 019XX N KENMORE AVE|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| fals
|11646447|   JC213946|2008-10-24 14:30:00|036XX N NARRAGANS...|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT| fals
|11646550|   JC214099|2018-10-04 12:00:00|   092XX S PERRY AVE|1120|  DECEPTIVE PRACTICE|             FORGERY|               OTHER| fals
| 4229528|   HL545852|2005-08-12 23:00:00|063XX S COTTAGE G...|3730|INTERFERENCE WITH...| OBSTRUCTING JUSTICE|           SIDEWALK|  tru
|11647549|   JC214237|2018-06-07 15:00:00|102XX S EGGLESTON...|0820|               THEFT|      $500 AND UNDER|               OTHER| fals
|11031104|   JA362043|2008-07-24 00:01:00| 031XX W FILLMORE ST|1563|         SEX OFFENSE|CRIMINAL SEXUAL A...|           APARTMENT| fals
|11648173|   JC216048|2018-11-01 00:00:00|062XX S MARSHFIEL...|1153|  DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           APARTMENT| fals
|11445072|   JB415614|2018-08-30 19:45:00| 031XX W HARRISON ST|2027|           NARCOTICS|         POSS: CRACK|POLICE FACILITY/V...|  tru
+--------+-----------+-------------------+--------------------+----+---------------------+--------------------+--------------------+-----
only showing top 20 rows
```

(04-05) Challenge questions

What is the most frequently reported non-criminal activity?

```
rc.show(5)
```

```
--+-----------------+--------------------+--------------------+------+--------+----+--------+----+--------------+--------+------------+
CR|      Primary Type|         Description|Location Description|Arrest|Domestic|Beat|District|Ward|Community Area|FBI Code|X Coordinate|
--+-----------------+--------------------+--------------------+------+--------+----+--------+----+--------------+--------+------------+
53|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| false|   false|0111|     001|  42|            32|      11|        null|
53|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| false|   false|0824|     008|  15|            63|      11|        null|
53|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|   false|2222|     022|  21|            71|      11|        null|
10|            THEFT|           OVER $500|           RESIDENCE| false|    true|0631|     006|   8|            44|      06|        null|
53|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|   false|2515|     025|  30|            19|      11|        null|
--+-----------------+--------------------+--------------------+------+--------+----+--------+----+--------------+--------+------------+
```

```
rc.select(col('Primary Type')).distinct().count()
```

```
36
```

```
rc.select(col('Primary Type')).distinct().show(35)
```

```
+--------------------+
|        Primary Type|
+--------------------+
|OFFENSE INVOLVING...|
|CRIMINAL SEXUAL A...|
|            STALKING|
|PUBLIC PEACE VIOL...|
|           OBSCENITY|
|               ARSON|
|            GAMBLING|
|   CRIMINAL TRESPASS|
|             ASSAULT|
|LIQUOR LAW VIOLATION|
| MOTOR VEHICLE THEFT|
|               THEFT|
|             BATTERY|
|             ROBBERY|
|            HOMICIDE|
|    PUBLIC INDECENCY|
| CRIM SEXUAL ASSAULT|
|   HUMAN TRAFFICKING|
|        INTIMIDATION|
|        PROSTITUTION|
|  DECEPTIVE PRACTICE|
|CONCEALED CARRY L...|
|         SEX OFFENSE|
|     CRIMINAL DAMAGE|
|           NARCOTICS|
|       OTHER OFFENSE|
|          KIDNAPPING|
|            BURGLARY|
```

```
|   WEAPONS VIOLATION|
|OTHER NARCOTIC VI...|
|INTERFERENCE WITH...|
|           RITUALISM|
|   DOMESTIC VIOLENCE|
|NON-CRIMINAL (SUB...|
|      NON - CRIMINAL|
+--------------------+
only showing top 35 rows
```

Double-click (or enter) to edit

```
rc.select(col('Primary Type')).distinct().orderBy(col('Primary Type')).show(36)
```

```
+--------------------+
|        Primary Type|
+--------------------+
|               ARSON|
|             ASSAULT|
|             BATTERY|
|            BURGLARY|
|CONCEALED CARRY L...|
| CRIM SEXUAL ASSAULT|
|     CRIMINAL DAMAGE|
|CRIMINAL SEXUAL A...|
|   CRIMINAL TRESPASS|
|  DECEPTIVE PRACTICE|
|   DOMESTIC VIOLENCE|
|            GAMBLING|
|            HOMICIDE|
|   HUMAN TRAFFICKING|
|INTERFERENCE WITH...|
|        INTIMIDATION|
|          KIDNAPPING|
|LIQUOR LAW VIOLATION|
| MOTOR VEHICLE THEFT|
|           NARCOTICS|
|      NON - CRIMINAL|
|        NON-CRIMINAL|
|NON-CRIMINAL (SUB...|
|           OBSCENITY|
|OFFENSE INVOLVING...|
|OTHER NARCOTIC VI...|
|       OTHER OFFENSE|
|        PROSTITUTION|
|    PUBLIC INDECENCY|
|PUBLIC PEACE VIOL...|
|           RITUALISM|
|             ROBBERY|
|         SEX OFFENSE|
|            STALKING|
|               THEFT|
|   WEAPONS VIOLATION|
+--------------------+
```

```
#In above we see three types of non-criminal activities
rc.select(col('Primary Type')).distinct().orderBy(col('Primary Type')).show(36,truncate=False)
```

```
+---------------------------------+
|Primary Type                     |
+---------------------------------+
|ARSON                            |
|ASSAULT                          |
|BATTERY                          |
|BURGLARY                         |
|CONCEALED CARRY LICENSE VIOLATION|
|CRIM SEXUAL ASSAULT              |
|CRIMINAL DAMAGE                  |
|CRIMINAL SEXUAL ASSAULT          |
|CRIMINAL TRESPASS                |
|DECEPTIVE PRACTICE               |
|DOMESTIC VIOLENCE                |
|GAMBLING                         |
|HOMICIDE                         |
|HUMAN TRAFFICKING                |
|INTERFERENCE WITH PUBLIC OFFICER |
|INTIMIDATION                     |
|KIDNAPPING                       |
|LIQUOR LAW VIOLATION             |
```

```
|MOTOR VEHICLE THEFT              |
|NARCOTICS                        |
|NON - CRIMINAL                   |
|NON-CRIMINAL                     |
|NON-CRIMINAL (SUBJECT SPECIFIED) |
|OBSCENITY                        |
|OFFENSE INVOLVING CHILDREN       |
|OTHER NARCOTIC VIOLATION         |
|OTHER OFFENSE                    |
|PROSTITUTION                     |
|PUBLIC INDECENCY                 |
|PUBLIC PEACE VIOLATION           |
|RITUALISM                        |
|ROBBERY                          |
|SEX OFFENSE                      |
|STALKING                         |
|THEFT                            |
|WEAPONS VIOLATION                |
+---------------------------------+
```

```
#creating new data frame to filter data
nc = rc.filter((col('Primary Type')=='NON - CRIMINAL') | (col('Primary Type')=='NON-CRIMINAL') | (col('Primary Type')=='NON-CRIMINAL (SUBJEC
nc.show(50)
```

```
+----+--------------------+--------------------+--------------------+------+--------+----+--------+----+--------------+--------+---------
|IUCR|        Primary Type|         Description|Location Description|Arrest|Domestic|Beat|District|Ward|Community Area|FBI Code|X Coordin
+----+--------------------+--------------------+--------------------+------+--------+----+--------+----+--------------+--------+---------
|5093|        NON-CRIMINAL|       LOST PASSPORT|AIRPORT TERMINAL ...|  true|   false|1651|     016|  41|            76|      26|     1100
|5093|        NON-CRIMINAL|       LOST PASSPORT|             TAXICAB| false|   false|0122|     001|  42|            32|      26|     1175
|5093|        NON-CRIMINAL|       LOST PASSPORT|            SIDEWALK| false|   false|1924|     019|  44|             6|      26|     1169
|5093|        NON-CRIMINAL|       LOST PASSPORT|              STREET| false|   false|1834|     018|  42|             8|      26|     1177
|5093|        NON-CRIMINAL|       LOST PASSPORT|COMMERCIAL / BUSI...| false|   false|1833|     018|  42|             8|      26|     1177
|5093|        NON-CRIMINAL|       LOST PASSPORT|              STREET| false|   false|1831|     018|  42|             8|      26|     1176
|5073|NON-CRIMINAL (SUB...|NOTIFICATION OF C...|           APARTMENT| false|    true|0931|     009|  16|            61|      26|     1163
|5073|NON-CRIMINAL (SUB...|NOTIFICATION OF C...|            SIDEWALK|  true|   false|2011|     020|  40|             2|      26|     1157
|5094|        NON-CRIMINAL|      FOUND PASSPORT|             CTA BUS| false|   false|0332|     003|   5|            43|      26|     1190
|5093|        NON-CRIMINAL|       LOST PASSPORT|           APARTMENT|  true|   false|1925|     019|  44|             6|      26|     1172
|5093|        NON-CRIMINAL|       LOST PASSPORT|            AIRCRAFT| false|   false|0813|     008|  23|            56|      26|     1145
|5114|      NON - CRIMINAL|   FOID - REVOCATION|POLICE FACILITY/V...| false|   false|0531|     005|   9|            50|      26|     1183
|5114|      NON - CRIMINAL|   FOID - REVOCATION|POLICE FACILITY/V...| false|   false|0434|     004|  10|            51|      26|     1192
|5093|        NON-CRIMINAL|       LOST PASSPORT|       BAR OR TAVERN| false|   false|1831|     018|  42|             8|      26|     1174
|5114|      NON - CRIMINAL|   FOID - REVOCATION|           RESIDENCE| false|   false|1221|     012|   1|            24|      26|     1163
|5114|      NON - CRIMINAL|   FOID - REVOCATION|GOVERNMENT BUILDI...| false|   false|1033|     010|  12|            30|      26|     1158
|5114|      NON - CRIMINAL|   FOID - REVOCATION|              STREET| false|   false|1622|     016|  38|            15|      26|     1133
|5073|NON-CRIMINAL (SUB...|NOTIFICATION OF C...|           APARTMENT| false|    true|0913|     009|  11|            60|      26|     1168
|5093|        NON-CRIMINAL|       LOST PASSPORT|            SIDEWALK| false|   false|1923|     019|  44|             6|      26|     1167
|5114|      NON - CRIMINAL|   FOID - REVOCATION|           RESIDENCE| false|   false|0831|     008|  15|            66|      26|     1158
|5114|      NON - CRIMINAL|   FOID - REVOCATION|PARKING LOT/GARAG...| false|   false|1034|     010|  25|            31|      26|     1162
|5114|      NON - CRIMINAL|   FOID - REVOCATION|           RESIDENCE| false|   false|0513|     005|  34|            49|      26|     1177
|5093|        NON-CRIMINAL|       LOST PASSPORT|       BAR OR TAVERN| false|   false|1925|     019|  44|             6|      26|     1170
|5114|      NON - CRIMINAL|   FOID - REVOCATION|POLICE FACILITY/V...| false|   false|0531|     005|   9|            50|      26|     1183
|5093|        NON-CRIMINAL|       LOST PASSPORT|               OTHER| false|   false|0113|     001|   2|            32|      26|     1177
|5114|      NON - CRIMINAL|   FOID - REVOCATION|RESIDENCE PORCH/H...| false|   false|0922|     009|  14|            58|      26|     1157
|0585|        NON-CRIMINAL|NOTIFICATION OF S...|            SIDEWALK| false|   false|1613|     016|  41|            10|      26|     1127
|5114|      NON - CRIMINAL|   FOID - REVOCATION|POLICE FACILITY/V...| false|   false|0823|     008|  15|            66|      26|     1154
|5093|        NON-CRIMINAL|       LOST PASSPORT|               OTHER| false|   false|2514|     025|  31|            19|      26|     1138
|5093|        NON-CRIMINAL|       LOST PASSPORT|        CTA PLATFORM| false|   false|1913|     019|  46|             3|      26|     1167
|5093|        NON-CRIMINAL|       LOST PASSPORT|           RESIDENCE| false|   false|2024|     020|  46|             3|      26|     1168
|5093|        NON-CRIMINAL|       LOST PASSPORT|              STREET| false|   false|1924|     019|  44|             6|      26|     1169
|5093|        NON-CRIMINAL|       LOST PASSPORT|           CTA TRAIN| false|   false|1623|     016|  45|            11|      26|     1139
|5093|        NON-CRIMINAL|       LOST PASSPORT|    FEDERAL BUILDING| false|   false|1911|     019|  47|             4|      26|     1161
|5093|        NON-CRIMINAL|       LOST PASSPORT|       PARK PROPERTY| false|   false|1834|     018|  42|             8|      26|     1180
|0585|        NON-CRIMINAL|NOTIFICATION OF S...|               ALLEY| false|   false|1031|     010|  22|            30|      26|     1149
|5114|      NON - CRIMINAL|   FOID - REVOCATION|           RESIDENCE| false|   false|0823|     008|  16|            66|      26|     1154
|5113|        NON-CRIMINAL|GUN OFFENDER NOTI...|            SIDEWALK| false|   false|0611|     006|  18|            71|      26|     1163
|5114|      NON - CRIMINAL|   FOID - REVOCATION|           RESIDENCE| false|   false|0832|     008|  18|            66|      26|     1160
|5093|        NON-CRIMINAL|       LOST PASSPORT|            SIDEWALK| false|   false|1834|     018|  42|             8|      26|     1177
|5093|        NON-CRIMINAL|       LOST PASSPORT|           RESIDENCE| false|   false|2423|     024|  49|             1|      26|     1163
|5093|        NON-CRIMINAL|       LOST PASSPORT|              STREET| false|   false|0112|     001|  42|            32|      26|     1176
|5094|        NON-CRIMINAL|      FOUND PASSPORT|           APARTMENT| false|   false|1033|     010|  12|            30|      26|     1156
|5114|      NON - CRIMINAL|   FOID - REVOCATION|POLICE FACILITY/V...| false|   false|1623|     016|  45|            11|      26|     1138
|5114|      NON - CRIMINAL|   FOID - REVOCATION|POLICE FACILITY/V...| false|   false|0213|     002|   3|            35|      26|     1177
|1481|        NON-CRIMINAL|CONCEALED CARRY L...|              STREET|  true|   false|0833|     008|  13|            65|      15|     1148
|5114|      NON - CRIMINAL|   FOID - REVOCATION|POLICE FACILITY/V...| false|   false|0225|     002|   3|            37|      26|     1175
|5114|      NON - CRIMINAL|   FOID - REVOCATION|           RESIDENCE| false|   false|0835|     008|  18|            70|      26|     1157
|5114|      NON - CRIMINAL|   FOID - REVOCATION|POLICE FACILITY/V...| false|   false|1414|     014|  35|            22|      26|     1157
|5114|      NON - CRIMINAL|   FOID - REVOCATION|           RESIDENCE| false|   false|0931|     009|  16|            63|      26|     1162
+----+--------------------+--------------------+--------------------+------+--------+----+--------+----+--------------+--------+---------
```

```
nc.groupBy(col('Description')).count().orderBy('count',ascending=False).show(truncate=False)
```

```
+-----------------------------------------+-----+
|Description                              |count|
+-----------------------------------------+-----+
|LOST PASSPORT                            |107  |
|FOID - REVOCATION                        |75   |
|NOTIFICATION OF CIVIL NO CONTACT ORDER   |9    |
|NOTIFICATION OF STALKING - NO CONTACT ORDER|8    |
|FOUND PASSPORT                           |4    |
|CONCEALED CARRY LICENSE REVOCATION       |4    |
|GUN OFFENDER NOTIFICATION-NO CONTACT     |3    |
+-----------------------------------------+-----+
```

Which day of the week has the most number of reported crime?

```
from pyspark.sql.functions import dayofweek
```

```
rc.show(5)
```

```
+--------+-----------+------------------+--------------------+----+-----------------+--------------------+--------------------+------+
|      ID|Case Number|              Date|               Block|IUCR|      Primary Type|         Description|Location Description|Arrest|
+--------+-----------+------------------+--------------------+----+-----------------+--------------------+--------------------+------+
|11037294|   JA371270|2015-03-18 12:00:00|   0000X W WACKER DR|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                BANK| false|
|11645836|   JC212333|2016-05-01 00:25:00| 055XX S ROCKWELL ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|                null| false|
|11645601|   JC212935|2014-06-01 00:01:00| 087XX S SANGAMON ST|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
|11646166|   JC213529|2018-09-01 00:01:00|082XX S INGLESIDE...|0810|            THEFT|           OVER $500|           RESIDENCE| false|
|11645648|   JC212959|2018-01-01 08:00:00| 024XX N MONITOR AVE|1153|DECEPTIVE PRACTICE|FINANCIAL IDENTIT...|           RESIDENCE| false|
+--------+-----------+------------------+--------------------+----+-----------------+--------------------+--------------------+------+
only showing top 5 rows
```

```
rc.select(col('Date'),dayofweek(col('Date'))).show(5)
```

```
+-------------------+---------------+
|               Date|dayofweek(Date)|
+-------------------+---------------+
|2015-03-18 12:00:00|              4|
|2016-05-01 00:25:00|              1|
|2014-06-01 00:01:00|              1|
|2018-09-01 00:01:00|              7|
|2018-01-01 08:00:00|              2|
+-------------------+---------------+
only showing top 5 rows
```

```
from pyspark.sql.functions import date_format
rc.select(col('Date'),dayofweek(col('Date')), date_format(col('Date'),'E')).show(5)
```

```
+-------------------+---------------+--------------------+
|               Date|dayofweek(Date)|date_format(Date, E)|
+-------------------+---------------+--------------------+
|2015-03-18 12:00:00|              4|                 Wed|
|2016-05-01 00:25:00|              1|                 Sun|
|2014-06-01 00:01:00|              1|                 Sun|
|2018-09-01 00:01:00|              7|                 Sat|
|2018-01-01 08:00:00|              2|                 Mon|
+-------------------+---------------+--------------------+
only showing top 5 rows
```

```
rc.groupBy(date_format(col('Date'),'E')).count().orderBy('count',ascending=False).show()
```

```
+--------------------+-------+
|date_format(Date, E)|  count|
+--------------------+-------+
|                 Fri|1017501|
|                 Wed| 974402|
|                 Tue| 968567|
|                 Sat| 965619|
|                 Thu| 965137|
|                 Mon| 953277|
|                 Sun| 912471|
```

```
+--------------------+-------+
```

```
rc.groupBy(date_format(col('Date'),'E')).count().collect()
```

```
[Row(date_format(Date, E)='Sun', count=912471),
 Row(date_format(Date, E)='Mon', count=953277),
 Row(date_format(Date, E)='Thu', count=965137),
 Row(date_format(Date, E)='Sat', count=965619),
 Row(date_format(Date, E)='Wed', count=974402),
 Row(date_format(Date, E)='Fri', count=1017501),
 Row(date_format(Date, E)='Tue', count=968567)]
```

Double-click (or enter) to edit

```
dow = [x[0] for x in rc.groupBy(date_format(col('Date'),'E')).count().collect()]
dow
```

```
['Sun', 'Mon', 'Thu', 'Sat', 'Wed', 'Fri', 'Tue']
```

```
cnt = [x[1] for x in rc.groupBy(date_format(col('Date'),'E')).count().collect()]
cnt
```

```
[912471, 953277, 965137, 965619, 974402, 1017501, 968567]
```

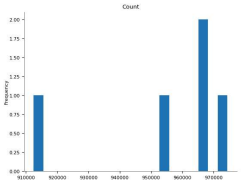
Using a bar chart, plot which day of the week has the most number of reported crime.

```
#lets plot this in graph
import pandas as pd
import matplotlib.pyplot as plt
```

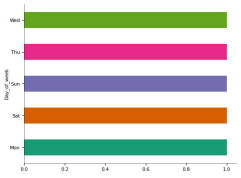
```
cp= pd.DataFrame({'Day_of_week': dow,'Count':cnt})
cp.head()
```

	Day_of_week	Count
0	Sun	912471
1	Mon	953277
2	Thu	965137
3	Sat	965619
4	Wed	974402

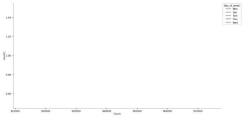
Distributions



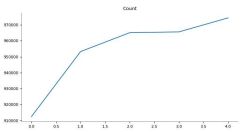
Categorical distributions



Time series



Values



Faceted distributions

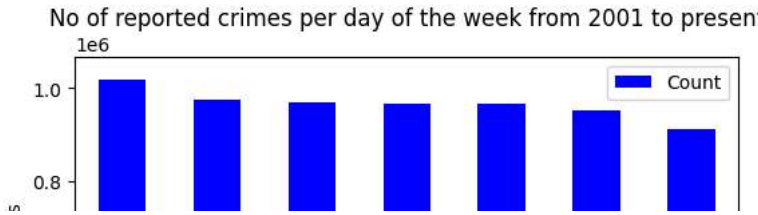
```
<string>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
```

Next steps: Generate code with cp View recommended plots

```
cp.sort_values('Count',ascending=False).plot(kind='bar', color ='blue',x='Day_of_week', y='Count')
plt.xlabel('Day of the week')
plt.ylabel('No of reported crimes')
plt.title('No of reported crimes per day of the week from 2001 to present')
```

```
Text(0.5, 1.0, 'No of reported crimes per day of the week from 2001 to present')
```



(05-01) RDDs setup

```
from pyspark import SparkContext

# Create a SparkContext object
sc = SparkContext.getOrCreate()
# Now you can use sc to perform operations
psrdd = sc.textFile('/content/police-stations.csv')
psrdd.first()
```

```
'DISTRICT,DISTRICT NAME,ADDRESS,CITY,STATE,ZIP,WEBSITE,PHONE,FAX,TTY,X COORDINATE,Y COO
RDINATE,LATITUDE,LONGITUDE,LOCATION'
```

```
ps_header = psrdd.first()


ps_rest = psrdd.filter(lambda line: line!= ps_header)
ps_rest.first()
```