POPU

Date 22/03/2021

Topic .....

Page No. 32

Experiment-10

Q. CSV file manipulation

1. Import file crime_incidents.csv into Databricks File System.

2. Import SQLContext

→ from pyspark.sql import SQLContext

3. Load the file and create Dataframe.

→ crimeincidentDF = spark.read.csv("dbfs:/FileStore/tables/crime_incidents.csv", inferSchema="true", header="true")

4. Collect

→ crimeincidentDF.collect()

5. Take.

→ crimeincidentDF.take(2)

6. Show

→ crimeincidentDF.show()

7. shows first 2 records in tabular format.

→ crimeincidentDF.show(2)

8. Finding no. of records.

→ crimeincidentDF.count()

Teacher's Signature .....

JAINA

9. Printing the first row of a data set.
→ crimeincidentDF.first()

10. Schema of dataframe.
→ crimeincidentDF.printSchema()

11. Describe dataframe.
→ crimeincidentDF.describe().show()

12. Select function.
→ newDF = crimeincidentDF.select('IncidntNum.', 'category', 'Descript')
newDF.show()

13. Selecting all columns.
→ newDF = crimeindidentDF.select("*")
newDF.show()

14. Correlation between 2 fields.
→ crimeincidentDF.corr('X', 'Y')

15. Remove Duplicate records.
→ newDF = crimeincidentDF.dropDuplicates()
newDF.show()

16. Finding name of a specific column using index.
→ crimeincident[4]

Teacher's Signature .....

JAINA

17 withColumn.

→
```
crimeincidentNC = crimeincidentDF.withColumn('IncidntNum', crimeincidentDF.IncidntNum).withColumn('category', crimeincidentDF.Category).withColumn('Resolution', crimeincidentDF.Resolution)
crimeincidentNC.show()
```

Q Registering table as a template, and applying SQL queries to it.

18 createOrReplaceTempView(table)

→
```
crimeincidentNC.createOrReplaceTempView("crime_incident_table")
newDF = sqlContext.sql("Select * from crime_incident_table WHERE Resolution LIKE \'ARRE%\'")
newDF.show()
```

19 Filtering Operations

→
```
subsetDF = crimeincidentNC.filter(crimeincidentsNC.Resolution == "NONE")
subsetDF.show(10)
```

20 Filtering through multiple conditions.

→
```
crimeincidentsNC.filter(crimeincidentsNC.Resolution == 'LOCATED').filter(crimeincidentsNC.Category == 'ASSAULT')
crimeincidentsNC.show()
```

Teacher's Signature

21. groupby

→ crimeincidentDF.groupBy(["category", crimeincidentDF.category]).count().show()

22. Applying Aggregation functions after grouping.

→ crimeincidentDF.groupby().avg().show()

23. Aggregation using column 'Resolution'

→ newDF = crimeincidentDF.groupBy(crimeincidentDF.Resolution)
newDF.count().show()

24. Display 15 elements of dataframe.

→ crimeincidentDF.show(15)

25. Writing the dataframe as Parquet file.

→ crimeincidentDF.write.parquet("dbfs:/FileStore/tables/crime_incidents.parquet")

26. The Parquet file will be stored by default in HDFS

→ newDF = spark.read.parquet("dbfs:/FileStore/tables/crime_incidents.parquet", inferSchema = "true", header = "true")
newDF.show()

Outputs:

Out[4]: [Row(IncidntNum = 50436712, Category = 'ASSAULT', Descript = 'BATTERY', DayOfWeek = 'Wednesday', Date = '04/20/2005 12:00:00 AM', Time = '04:00', PdDistrict = 'MISSION', Resolution = 'NONE', Address = '18th ST / CASTRO ST'),]

Out[6]:

IncidntNum	Category	Descript	DayOfWeek	Date	Time	PdDistrict
50436712	ASSAULT	BATTERY	Wednesday	04/20/15	04:00	MISSION
80049078	LARCENY/THEFT	GRAND THEFT	Sunday	01/13/8	18:00	PARK

Resolution	Address	X	Y	Location
NONE	18th ST / CASTRO	-122.4350	32.7460	(37.760...
NONE	1100 Block of	-122.4468	32.7462	(32.7625...

Out[8] crimeIncidentDF.count()
: 1888567

Out[9]: Row(IncidntNum = 50436712, Category = 'Assault', Description = 'BATTERY', DayOfWeek = 'Wednesday', Date = '04/20/2005 12:00:00 AM', Time = '04:00', PdDistrict = 'MISSION', Resolution = 'NONE', Address = '18TH ST / CASTRO ST', X = -122.4350, Y = 32.7608, Location = '(32.76091, -122.435...')

Teacher's Signature .....

Out[10]: root
 |-- IncidntNum: integer (nullable = true)
 |-- Category: string (nullable = true)
 |-- Descript: string (nullable = true)
 |-- DayOfWeek: string (nullable = true)
 |-- Date: string (nullable = true)
 |-- Time: string (nullable = true)
 |-- PdDistrict: string (nullable = true)
 |-- Resolution: string (nullable = true)
 |-- Address: string (nullable = true)
 |-- X: double (nullable = true)
 |-- Y: double (nullable = true)
 |-- Location: string (nullable = true)

Out[11]:

Summary	IncidntNum	Category	Descript	DayOfWeek	Date
count	1888567	1888562	1888562	1888562	1888562
mean	9.279556 39.	null	NULL	null	null
stddev	4.025831484.	null	NULL	null	null
min	- 3979	ARSON	CHILD	Friday	01/01/...
max	99158212 2.	WEAPON LAWS	COURT	Wednesday	12/31/...

Out[12]:

IncidntNum	Category	Descript
50436212	ASSAULT	BATTERY
80049078	LARCENY/THEFT	GRAND THEFT AUTO
130366639	ASSAULT	AGGRAVATED ASSAULT

Teacher's Signature .....

Out[14]: Correlation = 0.5565561354690178

Out[16]: crimeincidentOF[4]
Column <'Date'>

Out[18]:

IncidntNum	Category	Descript	DayOfWeek	Date	Time
130366639	ASSAULT	AGGRAVATED	SUNDAY	05/05/13	04:10
30810835	DRIVING	DRIVING	TUESDAY	7/8/3	12:00

PdDistrict	Resolution	Address	X	Y	Location
INGLESIDE	ARREST, BOOKED	0 Block of STCA	-122.444	32.72	(32.72...
SOUTHERN	ARREST, BOOKED	MASON ST	-122.408	37.78	(37.78...

Out[19]

IncidntNum	Category	Descript	DayOfWeek	Date	Time	PdDistrict
50436712	ASSAULT	BATTERY	Wednesday	4/20/5	04:00	MISSION
80049098	LARCENY	GRAND	Sunday	1/13/8	18:00	PARK

Resolution	Address	X	Y	Location
NONE	18th ST / CASTRO ST	-122.435...	37.760	(37.760,
NONE	1100 Block of CA...	-122.435...	37.762	(37.762,

Out[21]:

Category	Category	count
FRAUD	FRAUD	36014
SUICIDE	SUICIDE	1131
LIQUOR LAWS	LIQUOR LAWS	3860
SECONDARY CODES	SECONDARY CODES	21636

Teacher's Signature .....

Out[22]: Aggregation function

avg(IncidntNum)	avg(X)	avg(Y)
9.279556 3755 8	-122.427191 03	37.771759506

Out[23]:

Resolution	count
JUVENILE BOOKED	11789
EXCEPTIONAL CLEAR	3699
ARREST, BOOKED	445847
CLEARED-CONTACT	590

Out[25]:

path dbfs:/Filestore/tables/crime_incidents.parquet already exists.