Project A :

Chicago Crimes

Project Group 3

By
Alfredo Gonzalez
Nathan Lee
Nishitha Gouravelli

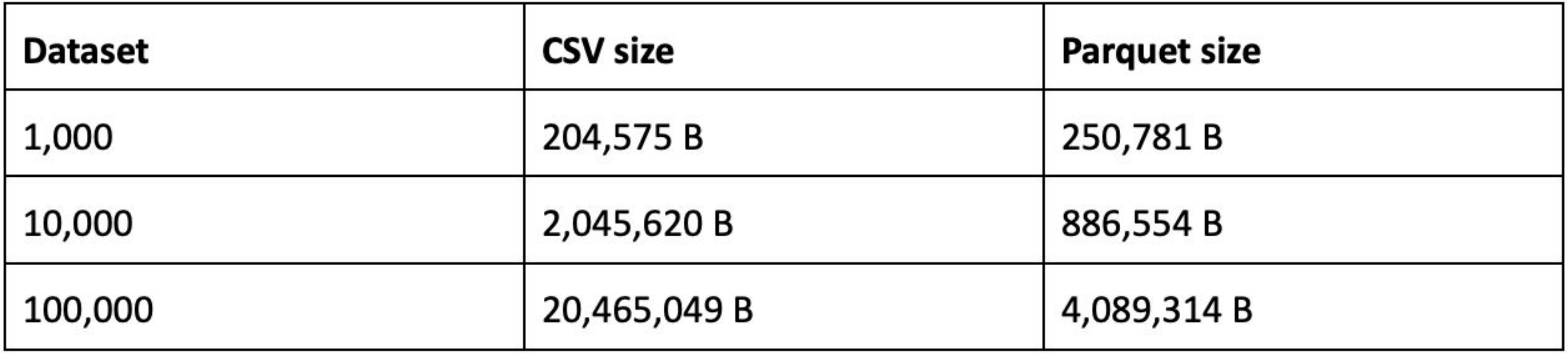
Summary

- This project analyzes Chicago Crimes using the dataset provided by the city of Chicago.
- The data set is first changed to make it easier to analyze and converted to Parquet file format.
- Then analysis is done to visualize the number of crimes for each ZIP code and the number of each type of crime committed between a time frame.

Task 1: Data preparation

- crimesDF.selectExpr("*", "ST_CreatePoint(x,y) AS geometry")
- RDD[(IFeature, IFeature)] = crimesRDD.spatialJoin(zipsRDD)
- crimesZipRDD.map({ case (crime, zip) => Feature.append(crime, zip.getAs[String]("ZCTA5CE10"), "ZIPCode") }).toDataFrame(sparkSession)
- finalDF.write.mode(SaveMode.Overwrite).parquet(outputFile)

Task 1: Data preparation



Dataset	CSV size	Parquet size
1,000	204,575 B	250,781 B
10,000	2,045,620 B	886,554 B
100,000	20,465,049 B	4,089,314 B

Task 2 : Spatial Analysis

- sparkSession.read.parquet(inputFile).createOrReplaceTempView(“<name>”)
- sparkContext.shapefile("tl_2018_us_zcta510.zip")

  .toDataFrame(sparkSession)

  .createOrReplaceTempView(“<name>”)

- sparkSession.sql(s"""

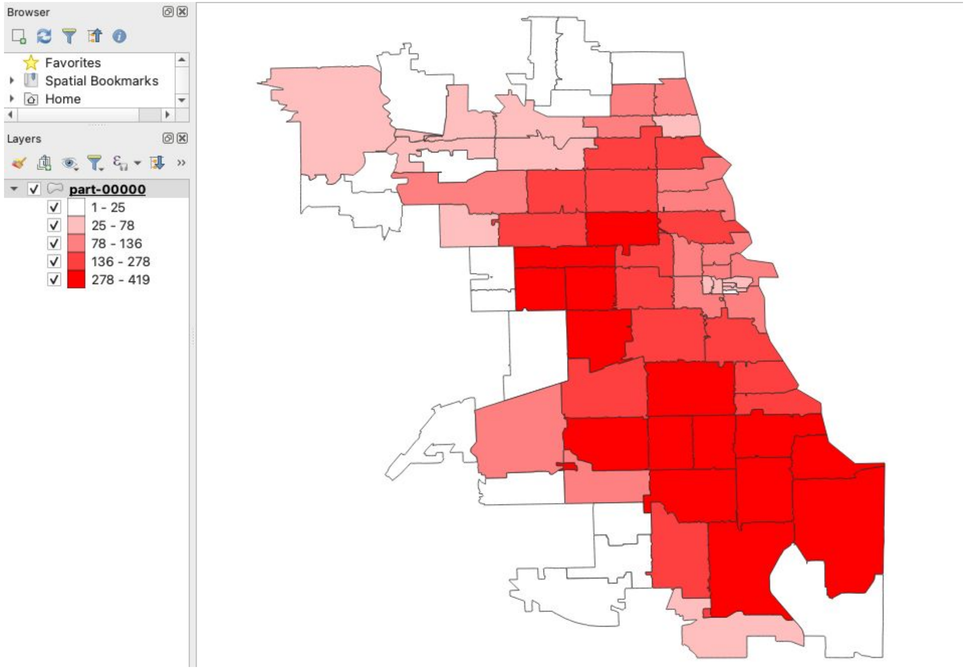
  ```
              // put queries here

          """)
  ```

- GROUP BY, count(*)
- WHERE ZIPCode = ZCTA5CE10
- .coalesce(1).saveAsShapefile("ZIPCodeCrimeCount")

Task 2 : Spatial Analysis

Visualization of the result for the 10k file from Task A :



Task 3 - Temporal Analysis

SparkSQL

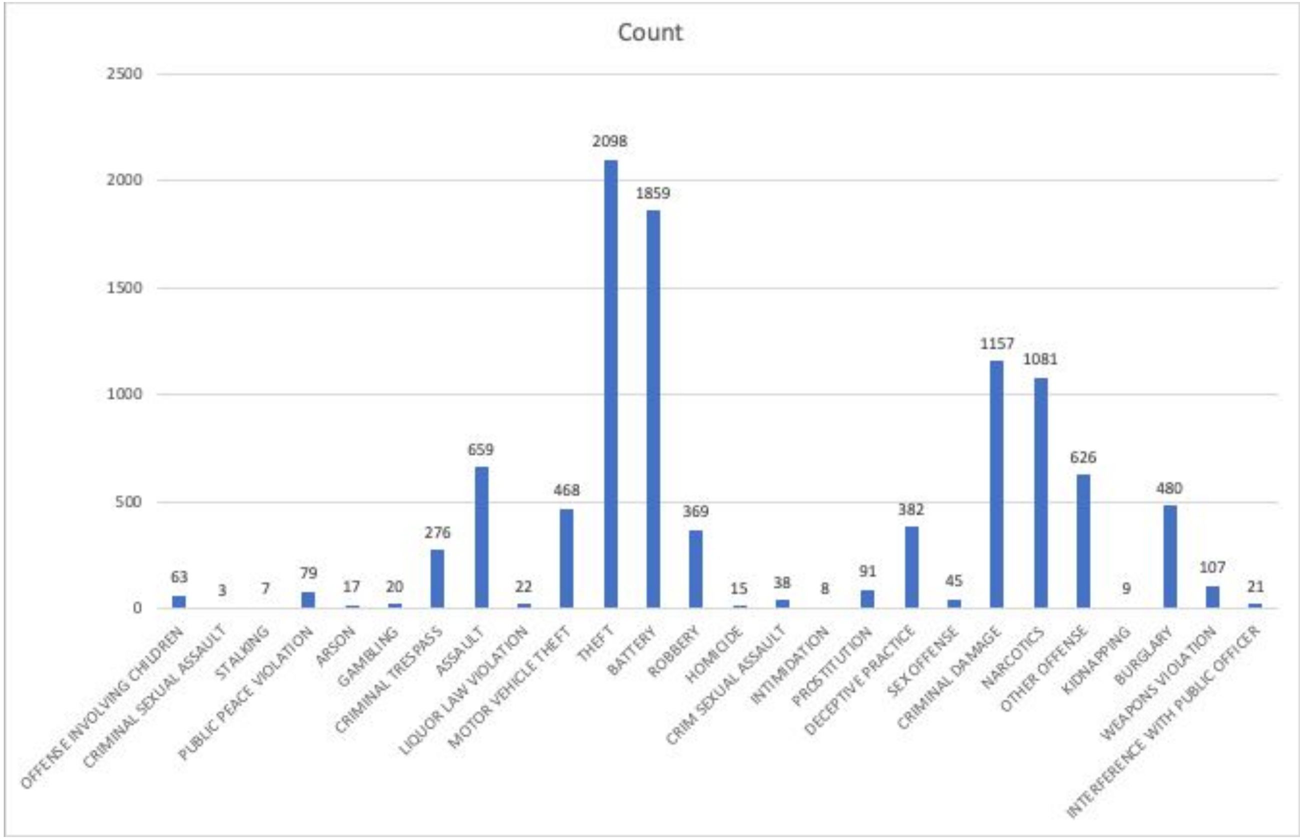
- read.parquet().createOrReplaceTempView()
- sql().coalsese(1).write.mode(SaveMode.Overwrite).option(header).csv()

SQL Functions

- to_timestamp - MM/dd/yyy hh:mm:ss a
- to_date - MM/DD/yyyy

Excel Bar Graph Function

Task 3 - Temporal Analysis



Thank you!