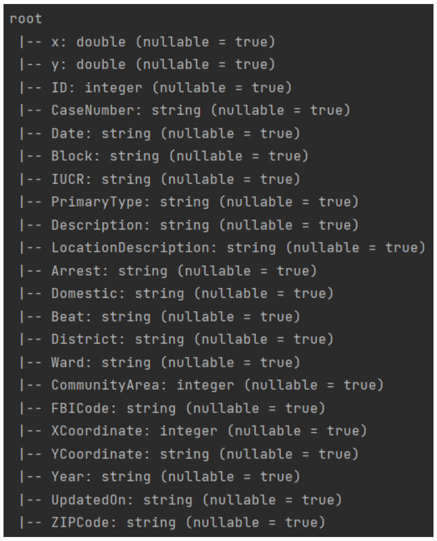
Hyunwook Yoo
Daniel Castaneda
Ricardo Villacana

Project A2 Report

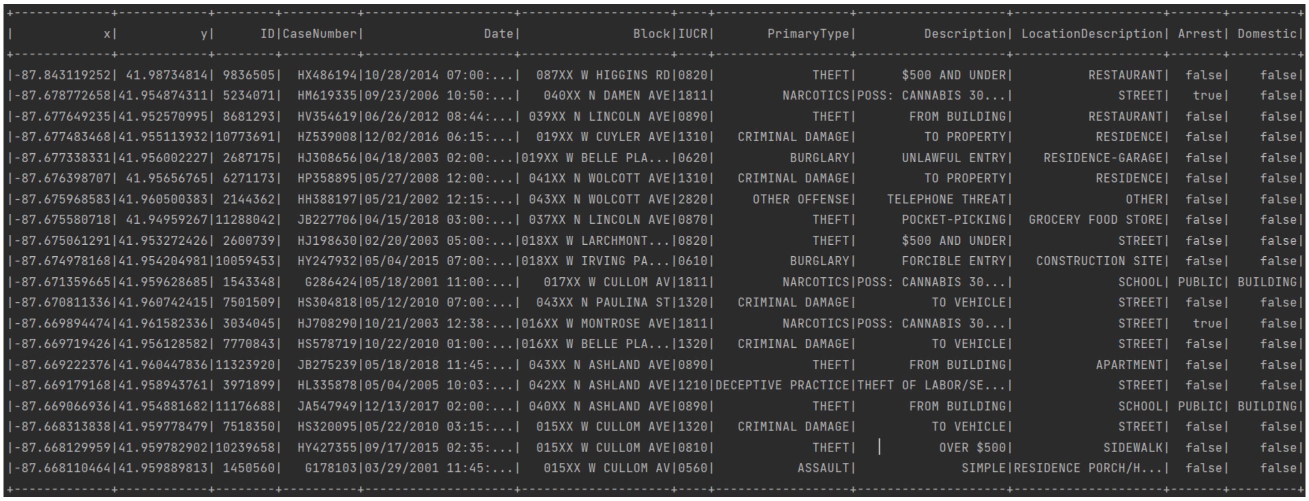
Task 1

In this task, we need to create one new column called ZIPCode to the crime dataset. To do that, we have to make a geometry column based on x column and y column from our crime dataset. After creating the geometry column, we run spatial join query with ZIP Code boundaries dataset and crime dataset with geometry column. And we create ZIPCode column to the crime dataset by using ZCTA5CE10 column from ZIPCode boundaries dataset and here is the schema for result:

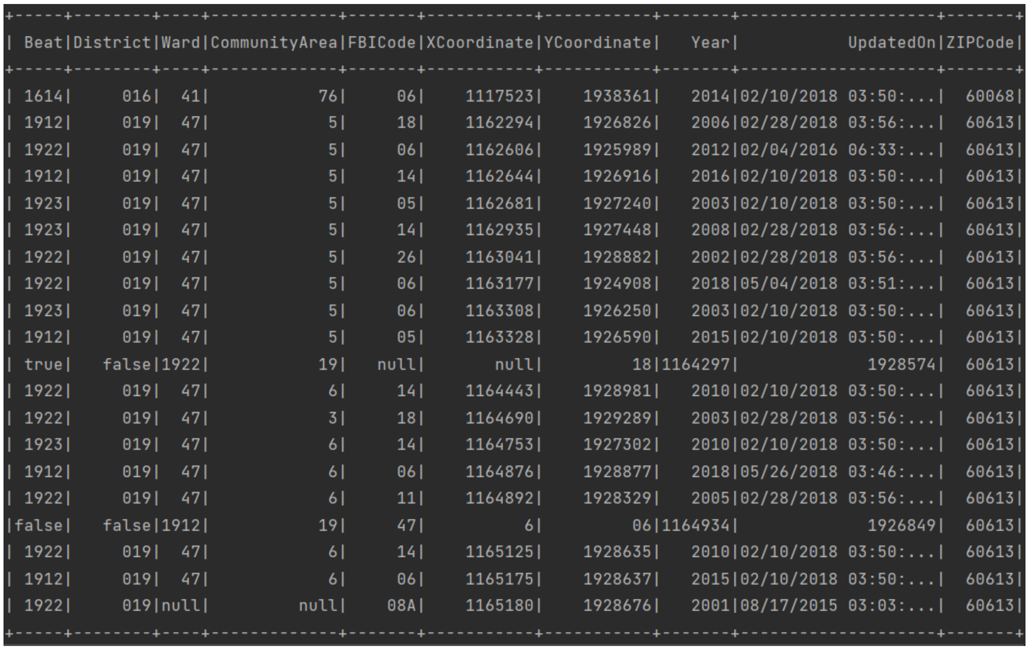


```
root
 |-- x: double (nullable = true)
 |-- y: double (nullable = true)
 |-- ID: integer (nullable = true)
 |-- CaseNumber: string (nullable = true)
 |-- Date: string (nullable = true)
 |-- Block: string (nullable = true)
 |-- IUCR: string (nullable = true)
 |-- PrimaryType: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- LocationDescription: string (nullable = true)
 |-- Arrest: string (nullable = true)
 |-- Domestic: string (nullable = true)
 |-- Beat: string (nullable = true)
 |-- District: string (nullable = true)
 |-- Ward: string (nullable = true)
 |-- CommunityArea: integer (nullable = true)
 |-- FBICode: string (nullable = true)
 |-- XCoordinate: integer (nullable = true)
 |-- YCoordinate: string (nullable = true)
 |-- Year: string (nullable = true)
 |-- UpdatedOn: string (nullable = true)
 |-- ZIPCode: string (nullable = true)
```

And this is the record of result dataset:



```
+-------------+------------+--------+----------+--------------------+--------------------+----+------------------+--------------------+--------------------+------+--------+
|            x|           y|      ID|CaseNumber|                Date|               Block|IUCR|       PrimaryType|         Description| LocationDescription|Arrest|Domestic|
+-------------+------------+--------+----------+--------------------+--------------------+----+------------------+--------------------+--------------------+------+--------+
|-87.843119252| 41.98734814| 9836505|  HX486194|10/28/2014 07:00:...|  087XX W HIGGINS RD|0820|             THEFT|      $500 AND UNDER|          RESTAURANT| false|   false|
|-87.678772658|41.954874311| 5234071|  HM619335|09/23/2006 10:50:...|   040XX N DAMEN AVE|1811|         NARCOTICS|POSS: CANNABIS 30...|              STREET|  true|   false|
|-87.677649235|41.952570995| 8681293|  HV354619|06/26/2012 08:44:...| 039XX N LINCOLN AVE|0890|             THEFT|       FROM BUILDING|          RESTAURANT| false|   false|
|-87.677483468|41.955113932|10773691|  HZ539008|12/02/2016 06:15:...|  019XX W CUYLER AVE|1310|   CRIMINAL DAMAGE|         TO PROPERTY|           RESIDENCE| false|   false|
|-87.677338331|41.956002227| 2687175|  HJ308656|04/18/2003 02:00:...|019XX W BELLE PLA...|0620|          BURGLARY|      UNLAWFUL ENTRY|    RESIDENCE-GARAGE| false|   false|
|-87.676398707| 41.95656765| 6271173|  HP358895|05/27/2008 12:00:...| 041XX N WOLCOTT AVE|1310|   CRIMINAL DAMAGE|         TO PROPERTY|           RESIDENCE| false|   false|
|-87.675968583|41.960500383| 2144362|  HH388197|05/21/2002 12:15:...| 043XX N WOLCOTT AVE|2820|     OTHER OFFENSE|    TELEPHONE THREAT|               OTHER| false|   false|
|-87.675580718| 41.94959267|11288042|  JB227706|04/15/2018 03:00:...| 037XX N LINCOLN AVE|0870|             THEFT|      POCKET-PICKING|  GROCERY FOOD STORE| false|   false|
|-87.675061291|41.953272426| 2600739|  HJ198630|02/20/2003 05:00:...|018XX W LARCHMONT...|0820|             THEFT|      $500 AND UNDER|              STREET| false|   false|
|-87.674978168|41.954204981|10059453|  HY247932|05/04/2015 07:00:...|018XX W IRVING PA...|0610|          BURGLARY|      FORCIBLE ENTRY|   CONSTRUCTION SITE| false|   false|
|-87.671359665|41.959628685| 1543348|   G286424|05/18/2001 11:00:...|   017XX W CULLOM AV|1811|         NARCOTICS|POSS: CANNABIS 30...|              SCHOOL|PUBLIC|BUILDING|
|-87.670811336|41.960742415| 7501509|  HS304818|05/12/2010 07:00:...|  043XX N PAULINA ST|1320|   CRIMINAL DAMAGE|          TO VEHICLE|              STREET| false|   false|
|-87.669894474|41.961582336| 3034045|  HJ708290|10/21/2003 12:38:...|016XX W MONTROSE AVE|1811|         NARCOTICS|POSS: CANNABIS 30...|              STREET|  true|   false|
|-87.669719426|41.956128582| 7770843|  HS578719|10/22/2010 01:00:...|016XX W BELLE PLA...|1320|   CRIMINAL DAMAGE|          TO VEHICLE|              STREET| false|   false|
|-87.669222376|41.960447836|11323920|  JB275239|05/18/2018 11:45:...| 043XX N ASHLAND AVE|0890|             THEFT|       FROM BUILDING|           APARTMENT| false|   false|
|-87.669179168|41.958943761| 3971899|  HL335878|05/04/2005 10:03:...| 042XX N ASHLAND AVE|1210|DECEPTIVE PRACTICE|THEFT OF LABOR/SE...|              STREET| false|   false|
|-87.669066936|41.954881682|11176688|  JA547949|12/13/2017 02:00:...| 040XX N ASHLAND AVE|0890|             THEFT|       FROM BUILDING|              SCHOOL|PUBLIC|BUILDING|
|-87.668313838|41.959778479| 7518350|  HS320095|05/22/2010 03:15:...|  015XX W CULLOM AVE|1320|   CRIMINAL DAMAGE|          TO VEHICLE|              STREET| false|   false|
|-87.668129959|41.959782902|10239658|  HY427355|09/17/2015 02:35:...|  015XX W CULLOM AVE|0810|             THEFT|           OVER $500|            SIDEWALK| false|   false|
|-87.668110464|41.959889813| 1450560|   G178103|03/29/2001 11:45:...|   015XX W CULLOM AV|0560|           ASSAULT|              SIMPLE|RESIDENCE PORCH/H...| false|   false|
+-------------+------------+--------+----------+--------------------+--------------------+----+------------------+--------------------+--------------------+------+--------+
```



```
+-----+--------+----+-------------+-------+-----------+-----------+-------+--------------------+-------+
| Beat|District|Ward|CommunityArea|FBICode|XCoordinate|YCoordinate|   Year|           UpdatedOn|ZIPCode|
+-----+--------+----+-------------+-------+-----------+-----------+-------+--------------------+-------+
| 1614|     016|  41|           76|     06|    1117523|    1938361|   2014|02/10/2018 03:50:...|  60068|
| 1912|     019|  47|            5|     18|    1162294|    1926826|   2006|02/28/2018 03:56:...|  60613|
| 1922|     019|  47|            5|     06|    1162606|    1925989|   2012|02/04/2016 06:33:...|  60613|
| 1912|     019|  47|            5|     14|    1162644|    1926916|   2016|02/10/2018 03:50:...|  60613|
| 1923|     019|  47|            5|     05|    1162681|    1927240|   2003|02/10/2018 03:50:...|  60613|
| 1923|     019|  47|            5|     14|    1162935|    1927448|   2008|02/28/2018 03:56:...|  60613|
| 1922|     019|  47|            5|     26|    1163041|    1928882|   2002|02/28/2018 03:56:...|  60613|
| 1922|     019|  47|            5|     06|    1163177|    1924908|   2018|05/04/2018 03:51:...|  60613|
| 1923|     019|  47|            5|     06|    1163308|    1926250|   2003|02/10/2018 03:50:...|  60613|
| 1912|     019|  47|            5|     05|    1163328|    1926590|   2015|02/10/2018 03:50:...|  60613|
| true|   false|1922|           19|   null|       null|         18|1164297|             1928574|  60613|
| 1922|     019|  47|            6|     14|    1164443|    1928981|   2010|02/10/2018 03:50:...|  60613|
| 1922|     019|  47|            3|     18|    1164690|    1929289|   2003|02/28/2018 03:56:...|  60613|
| 1923|     019|  47|            6|     14|    1164753|    1927302|   2010|02/10/2018 03:50:...|  60613|
| 1912|     019|  47|            6|     06|    1164876|    1928877|   2018|05/26/2018 03:46:...|  60613|
| 1922|     019|  47|            6|     11|    1164892|    1928329|   2005|02/28/2018 03:56:...|  60613|
|false|   false|1912|           19|     47|          6|         06|1164934|             1926849|  60613|
| 1922|     019|  47|            6|     14|    1165125|    1928635|   2010|02/10/2018 03:50:...|  60613|
| 1912|     019|  47|            6|     06|    1165175|    1928637|   2015|02/10/2018 03:50:...|  60613|
| 1922|     019|null|         null|    08A|    1165180|    1928676|   2001|08/17/2015 03:03:...|  60613|
+-----+--------+----+-------------+-------+-----------+-----------+-------+--------------------+-------+
```

The result dataset is Dataframe type, we need to create an output file as parquet type.

The reason that the parquet file is helpful for our project is that parquet format is column format for storage type. Column format is efficient for analytic query and is more efficient for compressing data than other storage format types. We can see how efficient for compressing the data when data size is bigger through table below:

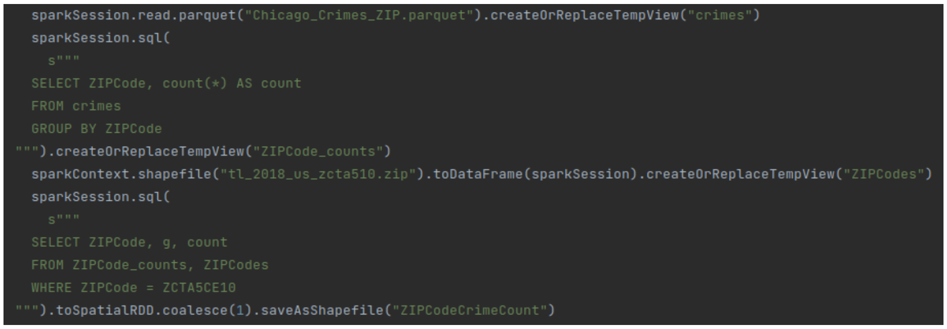
Dataset	CSV Size	Parquet Size
1,000	199 KB	242 KB
10,000	1.95 MB	858 KB
100,000	19.5 MB	7.11 MB

Like the table above, when the dataset is small, there is no difference between compressing the data. However, when data size is getting bigger, the efficiency of compressing data is higher. This is why parquet format is helpful for our project.

Task 2

The goal for this task is to do a spatial analysis for the data by counting the total number of crimes for each ZIP code and plotting the results as a choropleth map.

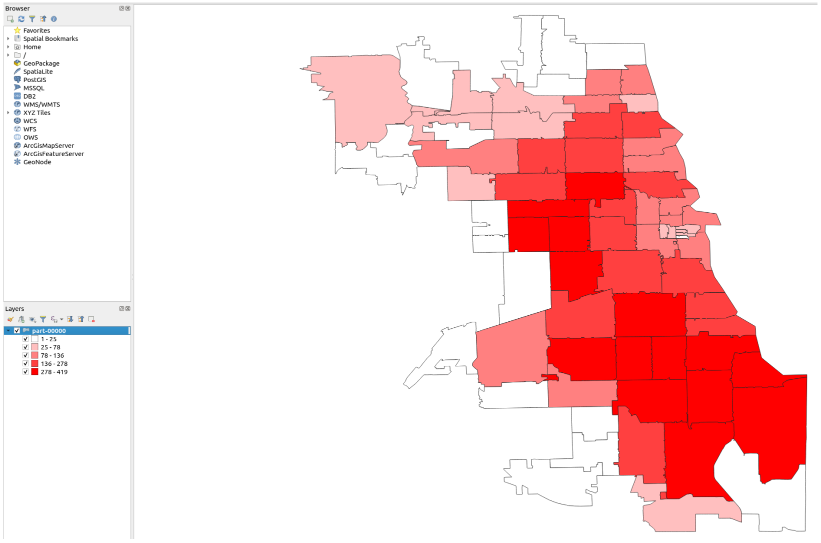
We start by loading the dataset in the Parquet format. Next, we need to run an SQL query that will compute the total number of crimes per ZIP code. In order to draw the choropleth map for these results, we need to join the results from the previous query with a ZIP code dataset. Finally, we save the results as a single shapefile which we will use to generate a choropleth map.



```
  sparkSession.read.parquet("Chicago_Crimes_ZIP.parquet").createOrReplaceTempView("crimes")
  sparkSession.sql(
    s"""
  SELECT ZIPCode, count(*) AS count
  FROM crimes
  GROUP BY ZIPCode
""").createOrReplaceTempView("ZIPCode_counts")
  sparkContext.shapefile("tl_2018_us_zcta510.zip").toDataFrame(sparkSession).createOrReplaceTempView("ZIPCodes")
  sparkSession.sql(
    s"""
  SELECT ZIPCode, g, count
  FROM ZIPCode_counts, ZIPCodes
  WHERE ZIPCode = ZCTA5CE10
""").toSpatialRDD.coalesce(1).saveAsShapefile("ZIPCodeCrimeCount")
```

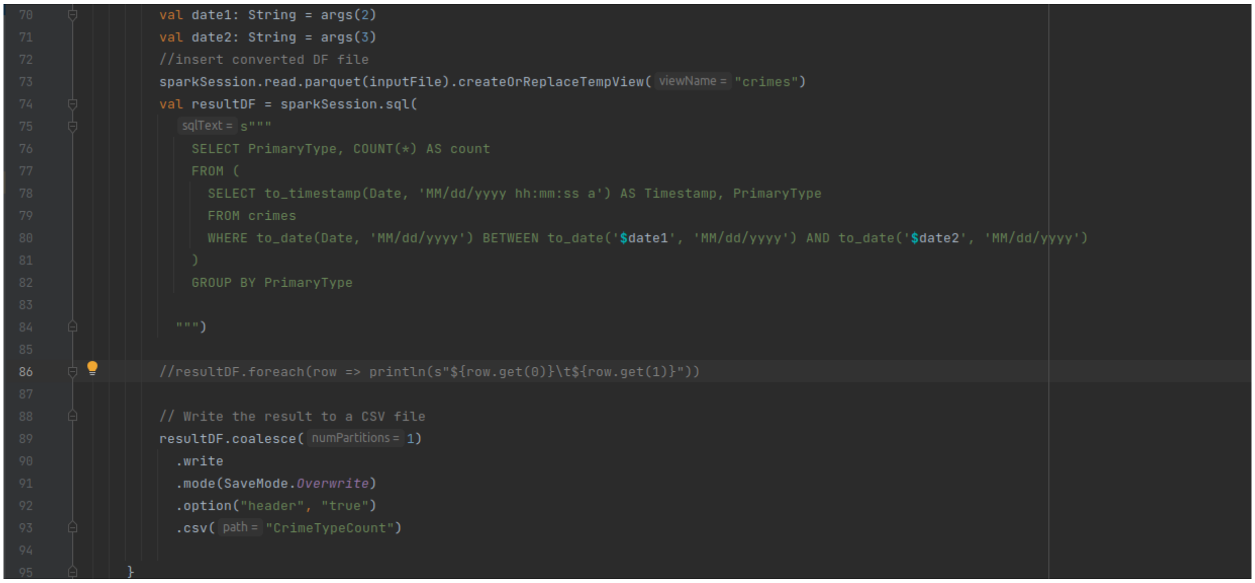
To generate a choropleth map, we use QGIS and pass in the .shp file that we just created. Once imported, we change to a graduated mode and set our value to count. Finally, we create the class with the new properties to apply our changes.

The following map is the resulting map for the 10k file:



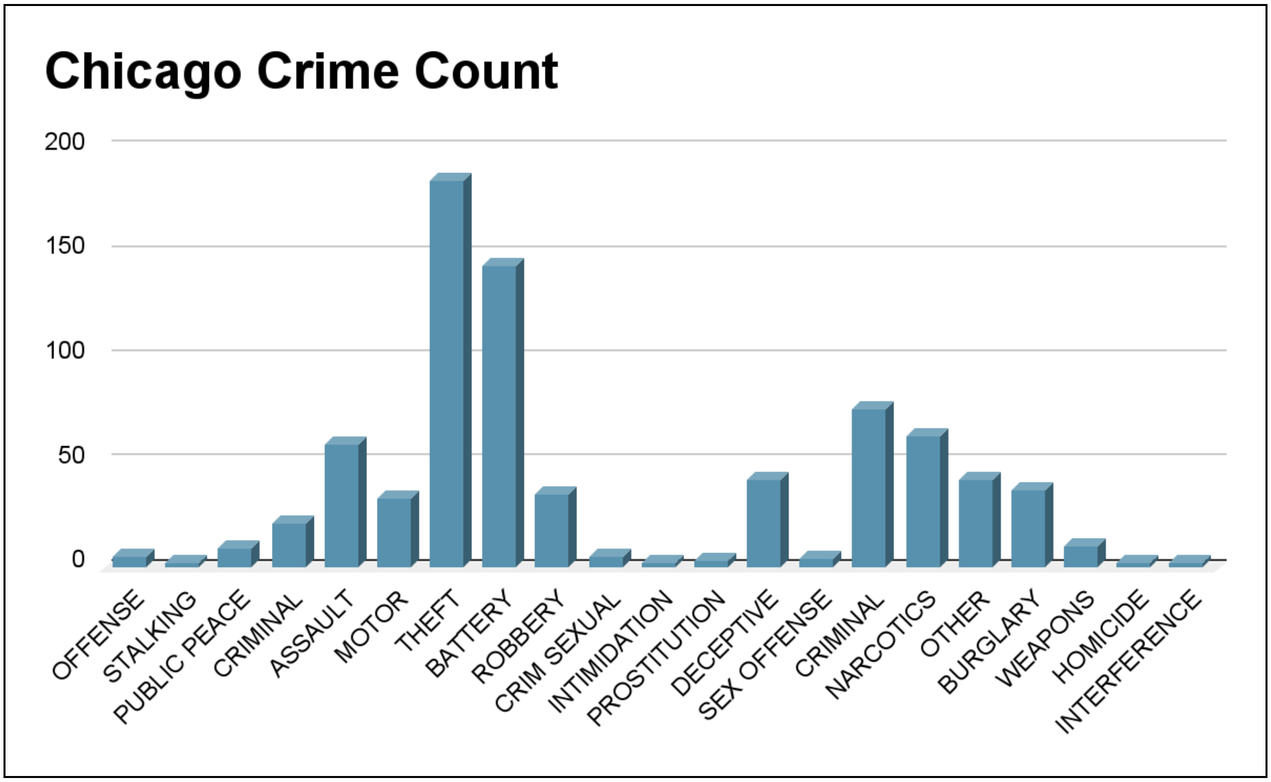
Task 3

The objective of task 3 is to use an sql query to count the number of crimes for each crime type in Chicago, given a start and end date, and the results are displayed with a histogram. The initial step is to take in the two date arguments, start date and end date. Then we load the converted dataset that is in parquet format, specifically the 10k dataset. After loading the dataset, we run a query that will count the number of crimes per crime type.



```
    val date1: String = args(2)
    val date2: String = args(3)
    //insert converted DF file
    sparkSession.read.parquet(inputFile).createOrReplaceTempView( viewName = "crimes")
    val resultDF = sparkSession.sql(
      sqlText = s"""
        SELECT PrimaryType, COUNT(*) AS count
        FROM (
          SELECT to_timestamp(Date, 'MM/dd/yyyy hh:mm:ss a') AS Timestamp, PrimaryType
          FROM crimes
          WHERE to_date(Date, 'MM/dd/yyyy') BETWEEN to_date('$date1', 'MM/dd/yyyy') AND to_date('$date2', 'MM/dd/yyyy')
        )
        GROUP BY PrimaryType

      """)

    //resultDF.foreach(row => println(s"${row.get(0)}\t${row.get(1)}"))

    // Write the result to a CSV file
    resultDF.coalesce( numPartitions = 1)
      .write
      .mode(SaveMode.Overwrite)
      .option("header", "true")
      .csv( path = "CrimeTypeCount")

  }
```

After running the query, we need to output the results to a csv file that only displays each crime type and the corresponding count for that crime type. The results in the csv file then need to be placed in and excel file to be represented with a histogram.



Chicago Crime Count