Chicago Crimes Report

Kimberly Morris, Neha Mathews, and Cynthia Wen

Computer Science 167

Professor Ahmed Eldawy

Group 11

June 8, 2022

Group 11 Chicago Crimes Project Tasks:

Task 1: Cynthia Wen

- Student ID: 861155604
- Email: cwen005@ucr.edu

Task 2: Kimbery Morris

- Student ID: 860896257
- Email: kmorr008@ucr.edu

Task 3: Neha Mathews

- Student ID: 862162878
- Email: nmath018@ucr.edu

Brief Introduction of Project:

In this project our group analyzed the Chicago Crimes dataset. The dataset contains over 14 million points, each with attributes about the crimes that occurred including Case Number, Type of crime, Location, etc. Our first task was to clean the data and convert to a parquet file to make data processing easier. We then created a choropleth map that showed the number of crimes per Zip Code within the city of Chicago. Lastly we created a bar chart that shows the number of each type of crime that occured within a specific date range.

Big Data System:

The big data system that we used for all three tasks was Spark SQL. The reason we opted to use this data system was because it was easier for us to write analytic queries such as finding the number of crimes per zip code and different types of crimes. Also, parquet files are recommended for analytical queries since it is column formatted.

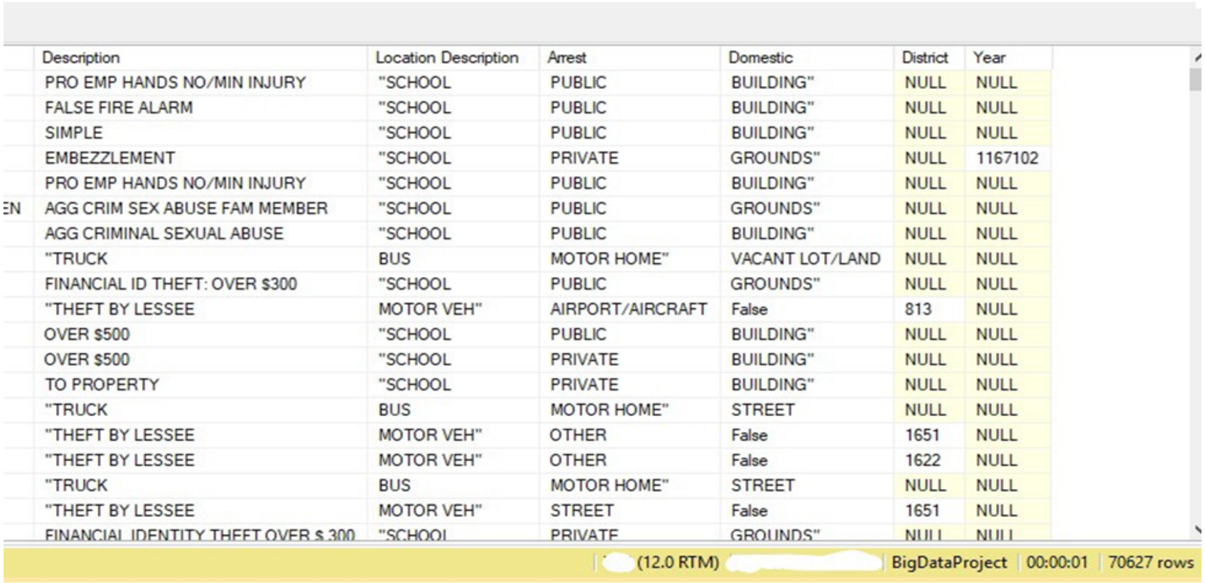
Introduction of Task 1:

According to the project's goals, the Chicago Crime dataset needs one of the most crucial data pre-processing procedures: cleaning. The initial step was cleaning the dataset to remove null values and null columns, the project dataset had a large section of it containing rows without the required number of columns to equate a record. Extracting just those attributes that are necessary for our data analysis (etc. X Coordinate, Y Coordinate, Latitude Longitude).

Also the dataset contained duplicated records that further reduced the integrity of the dataset, the dataset before pre- processing was so severely filled with invalid record the the size of each of the datasets reduced up to two thirds in size. For example, a file containing 1,923,865 entries was filtered to remove 70,627 records that did not match the column property. In order to make the data more accurate, I had to go through each column and delete any incorrect information. The duplicate data was then eliminated, and lastly, the column I didn't require was deleted in order for our data to be smaller and more quickly usable (see Figure). Also, any null values need to be deleted so that the genuine data may shine through. As a result, I execute a second query to locate all columns with null values. By doing this, the data was reduced from about 5 GB to 700 MB in size. All filter codings are located in the controller file.

After the clean up, Zip codes were loaded using Beast. Parquet files were created for 1k, 10k,100k datasets accordingly.

Parquet file is very much needed as it is successfully optimized to work with large and complex data. The Parquet format also did a great job on data compression in this case resulting in low storage consumption compared to csv format.



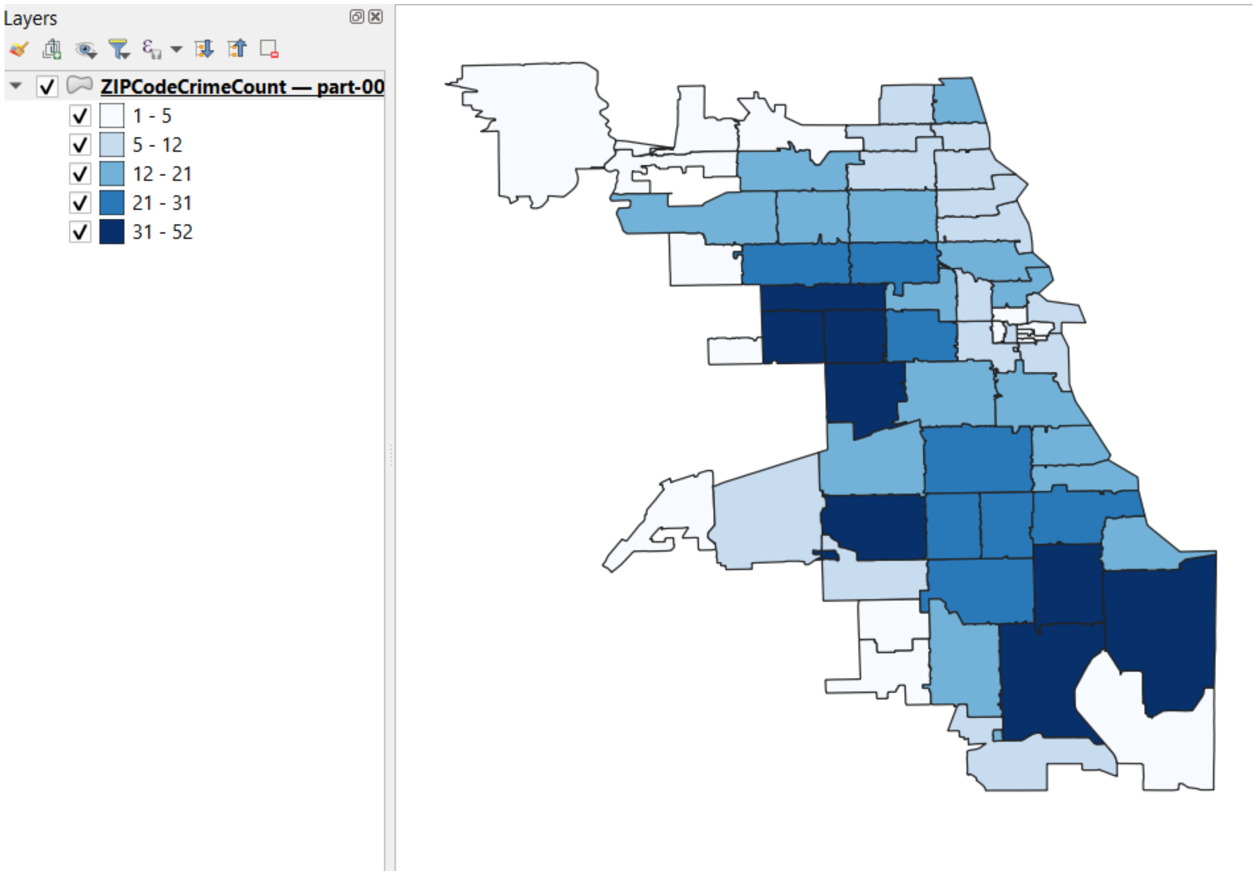
	Description	Location Description	Arrest	Domestic	District	Year
	PRO EMP HANDS NO/MIN INJURY	"SCHOOL	PUBLIC	BUILDING"	NULL	NULL
	FALSE FIRE ALARM	"SCHOOL	PUBLIC	BUILDING"	NULL	NULL
	SIMPLE	"SCHOOL	PUBLIC	BUILDING"	NULL	NULL
	EMBEZZLEMENT	"SCHOOL	PRIVATE	GROUNDS"	NULL	1167102
	PRO EMP HANDS NO/MIN INJURY	"SCHOOL	PUBLIC	BUILDING"	NULL	NULL
EN	AGG CRIM SEX ABUSE FAM MEMBER	"SCHOOL	PUBLIC	GROUNDS"	NULL	NULL
	AGG CRIMINAL SEXUAL ABUSE	"SCHOOL	PUBLIC	BUILDING"	NULL	NULL
	"TRUCK	BUS	MOTOR HOME"	VACANT LOT/LAND	NULL	NULL
	FINANCIAL ID THEFT: OVER $300	"SCHOOL	PUBLIC	GROUNDS"	NULL	NULL
	"THEFT BY LESSEE	MOTOR VEH"	AIRPORT/AIRCRAFT	False	813	NULL
	OVER $500	"SCHOOL	PUBLIC	BUILDING"	NULL	NULL
	OVER $500	"SCHOOL	PRIVATE	BUILDING"	NULL	NULL
	TO PROPERTY	"SCHOOL	PRIVATE	BUILDING"	NULL	NULL
	"TRUCK	BUS	MOTOR HOME"	STREET	NULL	NULL
	"THEFT BY LESSEE	MOTOR VEH"	OTHER	False	1651	NULL
	"THEFT BY LESSEE	MOTOR VEH"	OTHER	False	1622	NULL
	"TRUCK	BUS	MOTOR HOME"	STREET	NULL	NULL
	"THEFT BY LESSEE	MOTOR VEH"	STREET	False	1651	NULL
	FINANCIAL IDENTITY THEFT OVER $ 300	"SCHOOL	PRIVATE	GROUNDS"	NULL	NULL

[redacted] (12.0 RTM) [redacted] | BigDataProject | 00:00:01 | 70627 rows

DATASET	CSV SIZE	PARQUET SIZE
1,000	200 kb	94 kb
10,000	1998 kb	744 kb
100,000	19986 kb	6377 kb

Introduction of Task 2:

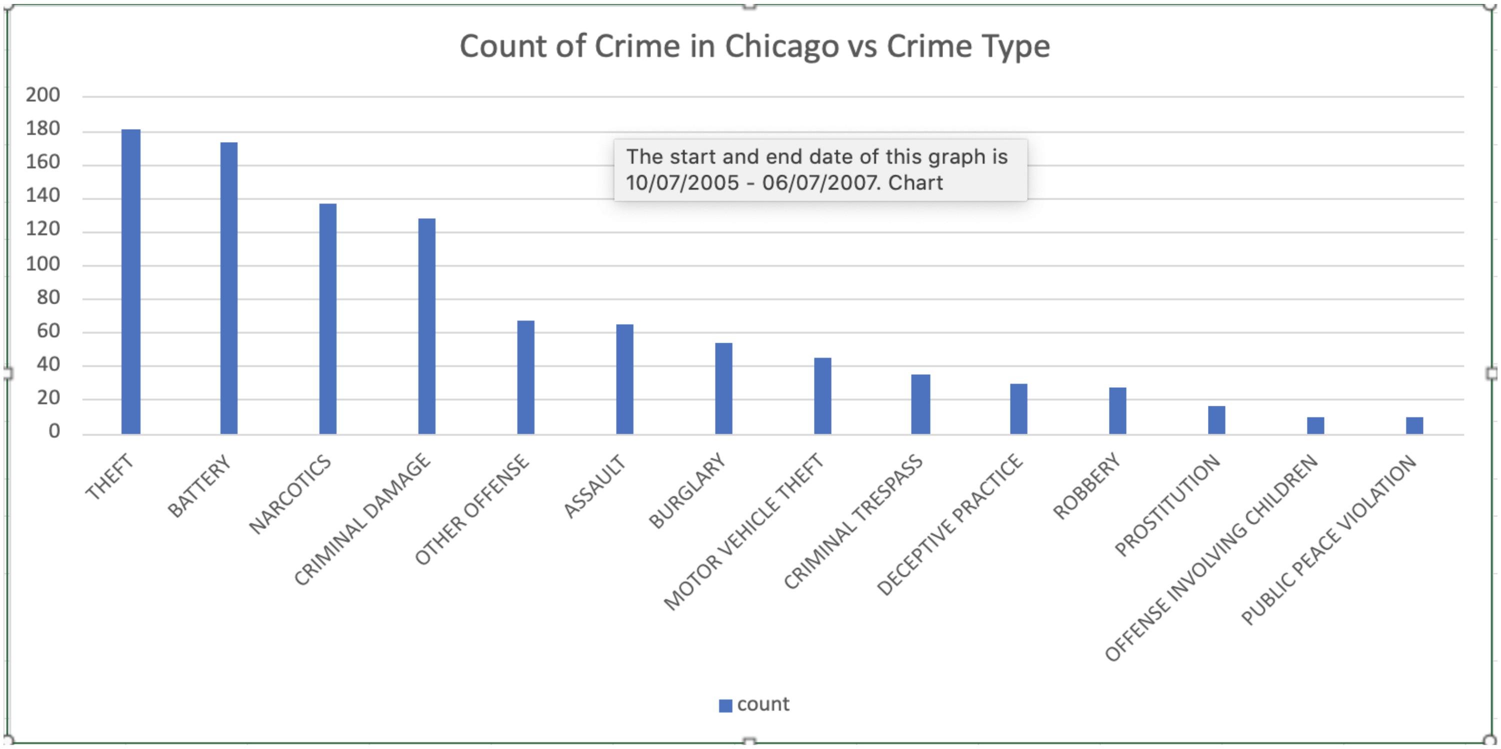
The goal of task Task 2 was to use the parquet file created in the step above to compute the total number of crimes in each Zip Code and plot the output on a choropleth map. To accomplish this goal I first created a view using an SQL query that selected the ZIPCode and count of all crimes grouped by ZIPCode. I then used Beast to load the ZIP Code dataset and convert it to a dataframe. Now that I had two views, I was able to use an equi-join query to join the two datasets on ZIPCode and save the output as a Shapefile. This output had the necessary geometry to create a choropleth map, and I imported the file into QGIS to produce the following map.



Introduction of Task 3:

For this task I was expected to create a temporal analysis of the collected chicago crimes data. I had to implement code that created a CSV file with the frequency of the primary type of the crime given a start and end date. Given this data, I had to create a bar chart of each of the crime types and its respective count. For my task I used Spark SQL. My program had two arguments which were the start and end date. I had to first read in the Chicago Crimes 10Kparquet file and create an output directory named, CrimeTypeCount. I used the function createOrreplaceTempView to have a temporary view of the data frame. In addition, I used printSchema() to print the schema and see what data types were included in the parquet file. I used an SQL query which usedto_ timestamp,to_date,WHERE,BETWEEN, and AND. Next I used the functions groupBy() and agg() which grouped all the crimes by Primary type and counted the frequency of each type of crime. Lastly, I used coalesce(1) to combine all the data frames into a single file before saving it in the directory. The output included a csv file which had the crimes from the start and end date. Using the csv file I was able to create a bar chart on excel as seen below.

Graph:



The start and end date of this graph is 10/07/2005 - 06/07/2007. Chart