Abstract

Crime analysis is a critical concern in urban management, particularly for cities like Kansas, where understanding crime dynamics directly impacts public safety and resource allocation. This project presents a detailed analysis of crime data spanning the years 2020 to 2024, focusing on identifying trends and patterns that influence crime rates and locations. By leveraging a robust preprocessing pipeline, which includes merging datasets, handling missing values, and standardizing critical features, we prepare the data for insightful analysis. We utilize Hive for efficient querying and exploration of crime distributions based on various factors such as time, type, and location, complemented by visualizations that elucidate key findings. The dataset encompasses a wide range of parameters, enabling the identification of crime hotspots and trends, thereby providing law enforcement and policymakers with actionable insights to enhance public safety measures. Through this analysis, we aim to equip stakeholders with data-driven strategies to effectively address crime and improve community safety in Kansas.

Abbreviations

HiveQL Hive Query Language

DVFlag Domestic Violence flag

IBRS Incident Based - *(classification system used in criminal justice data to categorize offenses)*

List of Figures

List of Visualisations

1. Introduction

Problem Statement:

To analyze and visualize the Crime data of the Kansas City in Missouri, USA.

Crime analysis is a crucial component in developing effective strategies for urban safety and resource management, helping law enforcement and policymakers make data-driven decisions to improve public welfare. In Kansas, the need for an in-depth understanding of crime trends and patterns has grown as the city seeks to address criminal activity efficiently and proactively. This project focuses on analyzing crime data from Kansas over the span of five years, from 2020 to 2024, aiming to identify patterns, trends, and potential correlations within the data. By examining both the temporal and spatial aspects of crime, this analysis provides insights that are valuable for public safety stakeholders and community leaders in Kansas.

To ensure data integrity and reliability, the project began with rigorous preprocessing of yearly crime datasets. The data preparation phase involved merging files from each year, renaming columns for clarity, handling missing values in key demographic and time variables, and standardizing features. Specific techniques, such as filling missing demographic data and converting date and time fields, were used to optimize the dataset for analysis. This comprehensive approach to data cleaning was crucial for enabling accurate analysis and setting a strong foundation for drawing meaningful insights. The resulting dataset captures key attributes like crime type, location, and timing, facilitating multi-dimensional analysis across the city.

Once the dataset was prepared, Hive was used to perform various analytical queries, allowing for efficient exploration of crime by offense type, time of day, and location, as well as demographic factors like age, race, and sex. This step was essential for identifying crime hotspots, seasonal trends, and other relevant patterns that may influence resource allocation and policy decisions. The findings were further enhanced through visualization, revealing critical insights that could help Kansas City officials prioritize interventions and deploy resources effectively. By delivering a comprehensive overview of crime trends in Kansas, this analysis supports efforts to improve public safety through data-driven decision-making and targeted crime prevention strategies.

2. Methodology

We have taken data of each year from the website and combined them into a single file.

Data-Link: - https://data.kcmo.org/browse?category=Crime

2.1 Dataset Description

The Kansas crime dataset now comprises **488,023 entries** and **31 columns**, offering an expansive view of crime incidents in Kansas City. Each row represents a unique crime report with detailed attributes capturing temporal, geographic, and demographic information. The key features in this dataset include:-

- **Report_No**: Unique identifier for each reported crime.
- **Reported_year, Reported_month, Reported_day, Reported_hour, Reported_minute**: Detailed breakdown of the date and time each crime was reported.
- **From_year, From_month, From_day, From_hour, From_minute**: Date and time when the crime was initially committed.
- **Offense**: Type of crime committed.
- **IBRS**: Incident-Based Reporting System code for the offense.
- **Description**: A detailed description of the offense.
- **Beat**: Numeric code representing the patrol beat for the incident.
- **Address**: Specific address of the incident.
- **City**: Specifies the city, consistently "KANSAS CITY" across all entries.
- **Zip_Code**: Zip code of the crime location.
- **Rep_Dist**: Reporting district within Kansas City.
- **Area**: Geographic area categorization for resource planning.
- **DVFlag**: Indicator of whether the incident involved domestic violence.
- **Involvement**: Type of involvement for individuals associated with the crime.
- **Race**: Race of the individuals involved.
- **Sex**: Gender of the individuals involved.
- **Age**: Age of individuals, with missing values imputed as needed.
- **Firearm_Used_Flag**: Boolean flag indicating firearm involvement.
- **Age_Group**: Categorical grouping of age for demographic analysis.
- **Age Range**: Range category for age, with some missing values.
- **Age Group**: Additional categorization of age for analytical flexibility.

- **Reported_Hour**: Hour at which the crime was reported, as a float.
- **Month_Year**: Concatenated month and year, allowing monthly trend analysis.

This dataset provides a valuable foundation for analyzing crime patterns in Kansas City over multiple years. The diversity of temporal, demographic, and geographic features enables thorough exploration of crime dynamics, supporting visualizations and insights into seasonal, spatial, and demographic trends.

```
RangeIndex: 488023 entries, 0 to 488022
Data columns (total 28 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Report_No       488023 non-null  object
 1   Reported_Date   488023 non-null  object
 2   Reported Time   96220 non-null   object
 3   From_Date       488017 non-null  object
 4   From Time       96219 non-null   object
 5   To_Date         230928 non-null  object
 6   To Time         32101 non-null   object
 7   Offense         488023 non-null  object
 8   IBRS            442699 non-null  object
 9   Description     442699 non-null  object
 10  Beat            488006 non-null  object
 11  Address         488023 non-null  object
 12  City            488020 non-null  object
 13  Zip Code        455154 non-null  object
 14  Rep_Dist        403560 non-null  object
 15  Area            488010 non-null  object
```

Report_No	Reported_year	Reported_month	Reported_day	Reported_hour	Reported_minute	From_yea	From_mo	From_day	From_hc	From_m	Offense	IBRS	Descriptio
KC21055624	2021	8	21	19	45	2021	8	21	19	45	Murder	09A	Murder
KC21012401	2021	2	24	11	35	2021	2	24	11	35	Stealing from Building/Residence	23D	Theft From
KC21010791	2021	2	17	11	41	2021	2	11	20	30	Stolen Auto	240	Motor Veh
KC21012025	2021	2	22	17	55	2021	2	22	17	55	Assault (Aggravated)	13A	Aggravate
KC21003742	2021	1	17	22	8	2021	1	17	21	30	Assault (Aggravated)	13A	Aggravate
KC21004380	2021	1	20	16	44	2021	1	20	16	44	Assault (Aggravated)	13A	Aggravate
KC21005430	2021	1	25	12	9	2020	4	13	12	0	Forgery	250	Counterfe
KC21008041	2021	2	5	11	14	2020	12	19	12	0	Identity Theft	26F	Identity Th
KC21011417	2021	2	20	1	41	2021	2	20	1	38	Robbery (Armed Street)	120	Robbery
KC21011702	2021	2	21	11	3	2021	2	20	17	0	Stolen Auto	240	Motor Veh
KC21012637	2021	2	25	11	45	2021	2	4	11	45	Identity Theft	26F	Identity Th
KC21010718	2021	2	17	0	6	2021	2	17	0	4	Murder	09A	Murder
KC21006497	2021	1	29	13	59	2021	1	29	15	45	Stealing â€“ Shoplift	23C	Shoplifting
KC21015438	2021	3	9	13	9	2021	3	9	13	9	Soliciting Prostitution	40A	Prostitutio
KC21009523	2021	2	11	11	27	2021	2	7	18	0	Stolen Auto	240	Motor Veh
KC21013729	2021	3	2	8	43	2021	3	1	21	30	Stealing from Auto (Theft from Aut	23F	Theft From
KC21009495	2021	2	11	8	45	2021	2	3	10	56	Embezzlement	270	Embezzler
KC21010861	2021	2	17	15	52	2021	2	17	15	52	Assault (Non-Aggravated)	13B	Simple Ass
KC21013274	2021	2	28	5	19	2021	2	28	5	19	Murder	09A	Murder
KC21009313	2021	2	10	14	1	2020	2	10	14	10	Stealing â€“ Shoplift	23C	Shoplifting
KC21006619	2021	1	30	2	57	2021	1	30	2	57	Murder	13A	Aggravate
KC21012281	2021	2	23	18	34	2021	2	23	18	30	Stolen Auto	240	Motor Veh
KC21011079	2021	2	18	16	41	2021	2	18	16	30	Patronizing Prostitution	40A	Prostitutio

Fig 1 Kansas Crime Data

In this project we used Hive and Pyspark to analyze and visualize the data.

2.2 Dataset loading

We loaded the data and had a quick overview of the data and checked for the null values. We got many Null value values, we modified the Null data points to some variable like any missing Sex is denoted as ‘U’ etc. We also renamed some columns for our easiness of use and merged all 5 data files.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 488023 entries, 0 to 488022
Data columns (total 28 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Report_No           488023 non-null  object
 1   Reported_Date       488023 non-null  object
 2   Reported Time       96220 non-null   object
 3   From_Date           488017 non-null  object
 4   From Time           96219 non-null   object
 5   To_Date             230928 non-null  object
 6   To Time             32101 non-null   object
 7   Offense             488023 non-null  object
 8   IBRS                442699 non-null  object
 9   Description         442699 non-null  object
 10  Beat                488006 non-null  object
 11  Address             488023 non-null  object
 12  City                488020 non-null  object
 13  Zip Code            455154 non-null  object
 14  Rep_Dist            403560 non-null  object
 15  Area                488010 non-null  object
 16  DVFlag              488023 non-null  object
 17  Involvement         488023 non-null  object
 18  Race                423268 non-null  object
 19  Sex                 429353 non-null  object
 20  Age                 360034 non-null  float64
 21  Firearm Used Flag   188347 non-null  object
 22  Location            445903 non-null  object
 23  Reported_Time       391803 non-null  object
 24  From_Time           391798 non-null  object
 25  To_Time             131281 non-null  object
 26  Age_Range           286877 non-null  object
 27  Fire Arm Used Flag  299676 non-null  object
dtypes: float64(1), object(27)
```

Fig 2 Data_info_before_processing

2.3 Preprocessing

Checked for null values

Missing values are addressed by filling in categorical variables like Sex and Race with 'U' for unknown, while the Age column is filled with the rounded mean age. Essential columns are filtered to exclude rows with null values, and a summary of remaining nulls is generated for transparency.

```
[91]: filtered_data.info()

<class 'pandas.core.frame.DataFrame'>
Index: 349353 entries, 0 to 349566
Data columns (total 31 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Report_No          349353 non-null  object
 1   Reported_year      349353 non-null  int64
 2   Reported_month     349353 non-null  int64
 3   Reported_day       349353 non-null  int64
 4   Reported_hour      349353 non-null  int64
 5   Reported_minute    349353 non-null  int64
 6   From_year          349353 non-null  int64
 7   From_month         349353 non-null  int64
 8   From_day           349353 non-null  int64
 9   From_hour          349353 non-null  int64
 10  From_minute        349353 non-null  int64
 11  Offense            349353 non-null  object
 12  IBRS               349353 non-null  object
 13  Description        349353 non-null  object
 14  Beat               349353 non-null  float64
 15  Address            349353 non-null  object
 16  City               349353 non-null  object
 17  Zip_Code           349353 non-null  object
 18  Rep_Dist           349353 non-null  object
 19  Area               349353 non-null  object
 20  DVFlag             349353 non-null  object
 21  Involvement        349353 non-null  object
 22  Race               349353 non-null  object
 23  Sex                349353 non-null  object
 24  Age                349353 non-null  float64
 25  Firearm_Used_Flag  349353 non-null  bool
 26  Age_Group          349353 non-null  category
 27  Age Range          347544 non-null  category
 28  Age Group          349353 non-null  category
 29  Reported_Hour      349353 non-null  float64
 30  Month_Year         349353 non-null  object
dtypes: bool(1), category(3), float64(3), int64(10), object(14)
```

Fig 3 Data_info_after_pre-processing

Date and time information is extracted into separate year, month, day, hour, and minute components, and unnecessary original columns are dropped. We made a column representing if the fire arms were used in the crime or not and another column representing if the crime was a domestic violence crime or not both were represented in true or false. We divided the ages into age groups giving a range of 10 for each age group.

3. Data analysis

We have used hive for data analysis. After preprocessing the complete data, we exported them into a csv file named ‘filtered_data.csv’ consisting of 31 columns and 349,353 rows in total.

3.1 Load data into Hadoop lfs

Create a new folder and then uploaded the file to the HDFS.

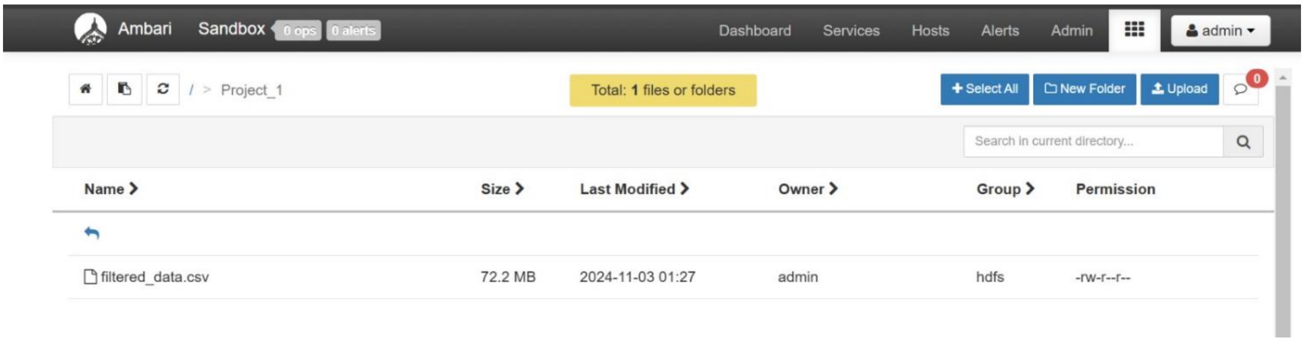
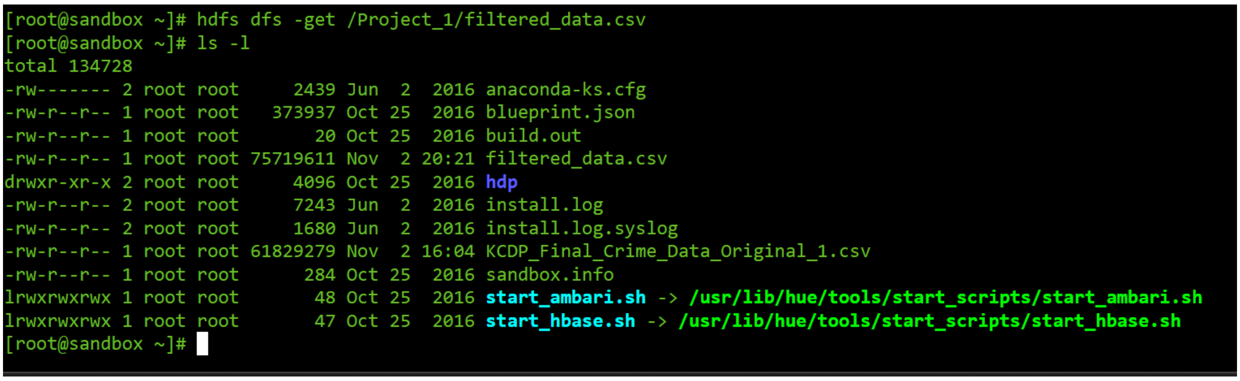


Fig 4 Hdfs File Loc

Get the function to the local file system in Hadoop.



```
[root@sandbox ~]# hdfs dfs -get /Project_1/filtered_data.csv
[root@sandbox ~]# ls -l
total 134728
-rw------- 2 root root     2439 Jun  2  2016 anaconda-ks.cfg
-rw-r--r-- 1 root root   373937 Oct 25  2016 blueprint.json
-rw-r--r-- 1 root root       20 Oct 25  2016 build.out
-rw-r--r-- 1 root root 75719611 Nov  2 20:21 filtered_data.csv
drwxr-xr-x 2 root root     4096 Oct 25  2016 hdp
-rw-r--r-- 2 root root     7243 Jun  2  2016 install.log
-rw-r--r-- 2 root root     1680 Jun  2  2016 install.log.syslog
-rw-r--r-- 1 root root 61829279 Nov  2 16:04 KCDP_Final_Crime_Data_Original_1.csv
-rw-r--r-- 1 root root      284 Oct 25  2016 sandbox.info
lrwxrwxrwx 1 root root       48 Oct 25  2016 start_ambari.sh -> /usr/lib/hue/tools/start_scripts/start_ambari.sh
lrwxrwxrwx 1 root root       47 Oct 25  2016 start_hbase.sh -> /usr/lib/hue/tools/start_scripts/start_hbase.sh
[root@sandbox ~]#
```

Fig 5 Loading file into LFS

3.2 Hive initiation

Creating a database and using that data base in hive.

```
hive> create database Crime_Database;
OK
Time taken: 0.157 seconds
hive> use Crime_Database;
OK
Time taken: 0.27 seconds
hive>
```

Creating the table

```
hive> create table crime_data(Report_No string,Reported_year int,Reported_month int,Reported_day int,Reported_hour int,Reported_minute int,F
rom_year int,From_month int,From_day int,From_hour int,From_minute int,offense string,IBRS string,Description string,Beat float,Address stri
ng,City string,Zip_code float,Rep_Dist string,Area string,DVFlag boolean,Involvement string,Race string,Sex string,Age float,Firearm_Used_Fl
ag boolean,Age_Group string,Age_Range string,Age_Group2 string,Reported_Hour_f float,Month_Year string)
    > row format delimited fields terminated by ',';
OK
Time taken: 0.631 seconds
hive>
```

Loading the data to the table

```
hive> load data local inpath 'filtered_data.csv' into table crime_data;
Loading data to table crime_database.crime_data
Table crime_database.crime_data stats: [numFiles=1, numRows=0, totalSize=75719611, rawDataSize=0]
OK
Time taken: 1.298 seconds
hive>
```

Verifying data from the table

```
hive> select * from crime_data limit 5;
OK
Report_No       NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    Offense IBRS    Description     NULL    Addr
ess     City    NULL    Rep_Dist        Area    NULL    Involvement     Race    Sex     NULL    NULL    Age_Group       Age Range       Age
Group   NULL    Month_Year
KC20017500      2020    3       8       19      24      2020    3       8       19      0       Stealing - Shoplift     90J     Trespass of
Real Property   345.0   11600   E US 40 HWY     KANSAS CITY     64133.0 PJ3601  EPD     false   VIC     U       U       38.0    false   35-4
4       30-40   30-39   19.4    3-2020
KC20009823      2020    2       7       18      30      2020    2       7       18      1       Robbery (Armed Street)  120     Robbery 542.
0       8100  E BANNISTER RD    KANSAS CITY     64134.0 PJ6554  SPD     false   SUS     B       M       38.0    true    35-44   30-40   30-3
9       18.5    2-2020
KC20011191      2020    2       13      12      15      2020    2       13      12      15      Stealing - Shoplift     23C     Shoplifting3
45.0    11600  E US 40 HWY      KANSAS CITY     64133.0 PJ3601  EPD     false   ARR CHA SUS     B       M       37.0    false   35-44   30-4
0       30-39   12.25   2-2020
KC20002142      2020    1       9       11      29      2020    1       8       13      50      Stealing - Other        23H     All Other La
rceny   414.0   5600  NW 78TH TER       KANSAS CITY     64151.0 PP0447  NPD     false   CMP VIC W       M       54.0    false   45-59   50-6
0       50-59   11.483334       1-2020
Time taken: 0.172 seconds, Fetched: 5 row(s)
hive>
```

Fig 6 Table in hive

select count(*) from crime_data;

```
hive> select count(*) from crime_data;
Query ID = root_20241102202730_66c6b4bb-57fb-4619-9361-e012e3f79b97
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0010)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 13.61 s
----------------------------------------------------------------------------------------------
OK
349354
Time taken: 15.456 seconds, Fetched: 1 row(s)
hive>
```

select distinct offense from crime_data;

```
hive> select distinct offense from crime_data;
Query ID = root_20241102202954_e177c751-c7c9-4c58-9c69-35a676427bff
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0010)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.74 s
----------------------------------------------------------------------------------------------
OK
Abandonment of a Child
Abuse of a Child
Adult Entertainment Violation
Alcohol Influence Report
Animal Abuse/Cruelty
Animal Bite
Arson
Assault (Aggravated)
Assault (Aggravated) on Department Member/Outside Law Enforcement Officer
Assault (Non-Aggravated)
Assault (Non-Aggravated) on Department Member/Outside Law Enforcement Officer
```

```
Stolen Auto
Suicide
Suicide - Attempted
Tampering
Tampering with Physical Evidence
Tavern/Nightclub Response Report
Terroristic Threats
Tobacco Law Violation
Tow-In Report/Authorization Not to Tow
Trafficking in Identifications
Trespass of Real Property
Unfounded
Unlawful Endangerment of Another
Unregistered Sex Offender
Vehicular - Fatality
Vehicular - Injury
Vehicular - Injury Hit and Run
Vehicular - Non-Injury
Vehicular - Non-Injury Hit and Run
Vehicular – Fatality
Vehicular – Injury
Vehicular – Injury Hit and Run
Vehicular – Non-Injury
Vehicular – Non-Injury Hit and Run
Violation of Ex-Parte Order of Protection
Violation of Full Order of Protection
Time taken: 12.344 seconds, Fetched: 156 row(s)
hive>
```

3.3 Data Analysis

1. **Yearly Trends over the years 2020 to 2024.**

```
select Reported_year,count(*) as total_crimes
from crime_data group by Reported_year
order by Reported_year;
```

```
hive> select Reported_year,count(*) as total_crimes
    > from crime_data
    > group by Reported_year
    > order by Reported_year;
Query ID = root_20241102205829_608ea929-3e09-4d9a-b2d7-f96187a499a7
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1730562569604_0011)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 12.76 s
----------------------------------------------------------------------------------
OK
NULL    1
2020    53703
2021    67425
2022    75414
2023    82141
2024    70670
Time taken: 18.13 seconds, Fetched: 6 row(s)
```

Fig 7 Query 1

We got 53703 crimes for 2020, 67425 crimes in the year 2021, 75414 for the year 2022 ,82141 crimes for the year 2023 and 70670 crimes for the year 2024.
We can observe that there is highest no of crimes int 2023.
The ranking of the no of crimes in the year goes like: -

1. 2023
2. 2022
3. 2024
4. 2021
5. 2020

2. **Analysis over the months**

```
select Reported_month,count(*) as total_crimes
from crime_data
group by Reported_month
order by Reported_month;
```

```
hive> select Reported_month,count(*) as total_crimes
    > from crime_data
    > group by Reported_month
    > order by Reported_month;
Query ID = root_20241102210108_64869534-5c3e-40de-9ce1-5af062d9a192
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0011)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 12.43 s
----------------------------------------------------------------------------------------------
OK
NULL    1
1       28181
2       25768
3       28173
4       28681
5       31411
6       32557
7       34625
8       34885
9       33141
10      29380
11      22644
12      19907
Time taken: 13.119 seconds, Fetched: 13 row(s)
hive>
```

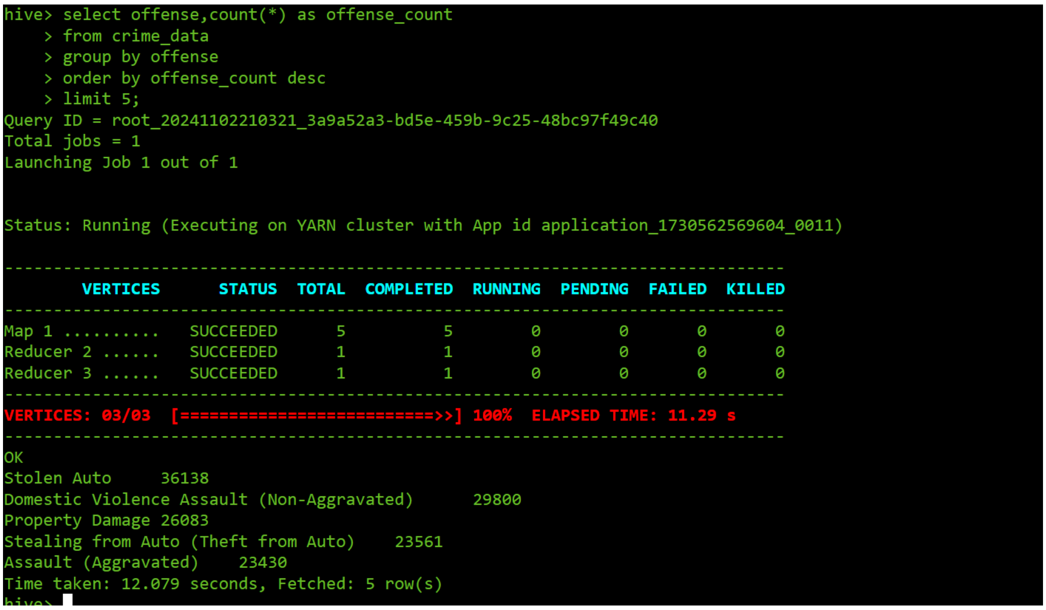
Fig 8 Query 2

Over the Period of 12 months across the year 2020 to 2025 , we have a ranking of : -

1. August: 34885
2. July: 34625
3. September: 33141
4. June: 32557
5. May: 31411
6. October: 29380
7. April: 28681
8. January: 28181
9. March: 28173
10. February: 25768
11. November: 22644
12. December: 19907

3. **Top 5 offenses in the data: -**

```
select offense,count(*) as offense_count
from crime_data
group by offense
order by offense_count desc
limit 5;
```



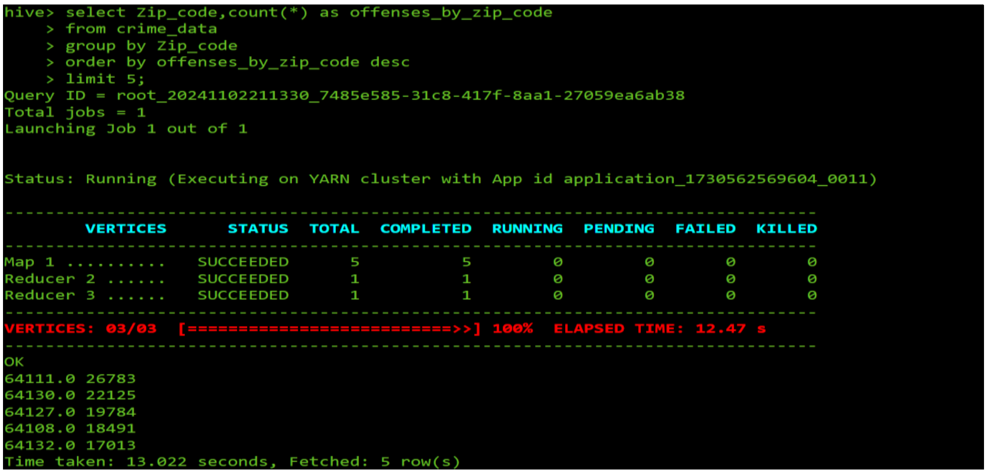
```
hive> select offense,count(*) as offense_count
    > from crime_data
    > group by offense
    > order by offense_count desc
    > limit 5;
Query ID = root_20241102210321_3a9a52a3-bd5e-459b-9c25-48bc97f49c40
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0011)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.29 s
----------------------------------------------------------------------------------
OK
Stolen Auto     36138
Domestic Violence Assault (Non-Aggravated)      29800
Property Damage 26083
Stealing from Auto (Theft from Auto)    23561
Assault (Aggravated)    23430
Time taken: 12.079 seconds, Fetched: 5 row(s)
hive>
```

Fig 9 Query 3

We examined the frequency of various offenses recorded in the crime dataset. By aggregating the data, we identified the top five most prevalent offenses, highlighting the most common types of criminal activity reported. This insight not only sheds light on the current crime landscape but also aids in understanding trends and patterns within the community.

4. Top 5 Offenses by ZIP Code

```
select Zip_code,count(*) as offenses_by_zip_code
from crime_data
group by Zip_code
order by offenses_by_zip_code desc
limit 5;
```



```
hive> select Zip_code,count(*) as offenses_by_zip_code
    > from crime_data
    > group by Zip_code
    > order by offenses_by_zip_code desc
    > limit 5;
Query ID = root_20241102211330_7485e585-31c8-417f-8aa1-27059ea6ab38
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0011)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 12.47 s
----------------------------------------------------------------------------------------------
OK
64111.0 26783
64130.0 22125
64127.0 19784
64108.0 18491
64132.0 17013
Time taken: 13.022 seconds, Fetched: 5 row(s)
```

Fig 10 Query 4

This query retrieves the top five ZIP codes with the highest number of reported offenses, providing insights into geographic crime distribution. It counts the total offenses for each ZIP code and sorts the results in descending order.

5. **Average Age of Offenders by Offense Type**

```
select offense,avg(Age) as Avg_age_offenders
from crime_data
group by offense
order by Avg_age_offenders desc;
```

```
hive> select offense,avg(Age) as Avg_age_offenders
    > from crime_data
    > group by offense
    > order by Avg_age_offenders desc;
Query ID = root_20241102211805_6e1e6675-673e-48a9-9107-c1bfd0dc0bba
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0011)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.36 s
----------------------------------------------------------------------------------------------
OK
Possession of Gambling Device or Records        71.0
Financial Exploitation of the Elderly   64.11428571428571
Elder Abuse     54.734939759036145
Curfew Violation        48.4
Animal Bite     47.0
Identity Theft  44.062875309078066
Domestic Violence Lethality Screen for First Responders (LAP)   44.0
```

```
Abandonment of a Child  34.529411764705884
Rape - Statutory        34.493506493506494
Kidnapping      34.46712802768166
Invasion of Privacy     34.16564417177914
Murder  34.14713656387665
Eluding / Resisting a Lawful Stop       34.01764057331863
Adult Entertainment Violation   34.0
Domestic Violence Burglary (Residential)        33.857290589451914
False Imprisonment      33.6
Domestic Violence Robbery (Strong-Armed)        33.536986301369865
Liquor Law Violation    33.526315789473685
Cold Case Sex Offense   33.25
Rape - Statutory        33.11904761904762
Suicide - Attempted     32.75
Police Vehicle Damage (Form 154 P.D.)   32.4
Human Trafficking/Commercial Sex Acts   32.25581395348837
Outside Correspondence  32.0
Escape from Custody/Confinement 32.0
Officer Involved Shooting - Fatal       31.695652173913043
Tobacco Law Violation   31.666666666666668
Tavern/Nightclub Response Report        31.166666666666668
Interdiction    30.75
Possession of Illegal Firearm   30.37074829931973
Vehicular - Fatality    30.0
Suicide 27.0
Offense NULL
Time taken: 13.993 seconds, Fetched: 156 row(s)
hive>
```

Fig 11 Query 5

This query calculates the average age of offenders for each type of offense. By grouping the data by offense and ordering the results in descending order, it highlights which offenses are associated with older or younger offenders, aiding in demographic analysis.

6. **Firearm count**

```
select count(case when Firearm_Used_Flag = True then 1 end) as
Firemarm_Used,
count(case when Firearm_Used_Flag = False then 1 end) as
Firearm_not_used
from crime_data;
```

```
hive> select count(case when Firearm_Used_Flag = True then 1 end) as Firemarm_Used,
    > count(case when Firearm_Used_Flag = False then 1 end) as Firearm_not_used
    > from crime_data;
Query ID = root_20241102214929_77dfeb0d-0694-405a-91fc-6391facede2b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 .........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......  SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.81 s
----------------------------------------------------------------------------------
OK
31599   315841
Time taken: 15.535 seconds, Fetched: 1 row(s)
hive>
```

Fig 12 Query 6

This query counts the number of offenses involving firearms versus those that do not. By separating the counts based on the `Firearm_Used_Flag`, it provides insights into the prevalence of firearm use in crimes, which is essential for understanding safety and policy implications.

7. **DvFlag Count**

```
select count(case when DVFlag = True then 1 end) as
Domestic_violence,
count(case when DVFlag = False then 1 end) as
Non_Domestic_violence
from crime_data;
```

```
hive> select count(case when DVFlag = True then 1 end) as Domestic_violence,
    > count(case when DVFlag = False then 1 end) as Non_Domestic_violence
    > from crime_data;
Query ID = root_20241102215402_e184be19-e377-4c79-aee1-a1d47334c63a
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.87 s
----------------------------------------------------------------------------------------------
OK
31580   315860
Time taken: 12.409 seconds, Fetched: 1 row(s)
hive>
```

Fig 13 Query 7

This query distinguishes between domestic violence and non-domestic violence incidents within the dataset. By counting occurrences based on the `DVFlag`, it offers valuable insights into the prevalence of domestic violence, aiding in resource allocation and community safety strategies.

8. **Firearm Usage by Offense Type**

```
select offense,count(*) as Firearm_Usage
from crime_data
where Firearm_Used_Flag = True
group by offense
order by Firearm_Usage desc;
```

```
hive> select offense,count(*) as Firearm_Usage
    > from crime_data
    > where Firearm_Used_Flag = True
    > group by offense
    > order by Firearm_Usage desc;
Query ID = root_20241102215645_3c24bdce-0f0c-4c31-9e2d-8e5cb7c1134f
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.81 s
----------------------------------------------------------------------------------------------
OK
Assault (Aggravated)    15133
Robbery (Armed Street)  4388
Domestic Violence Assault (Aggravated)  3361
Murder  1993
Robbery (Business)      1506
City Weapons Offense    1270
```

```
Stealing - Shoplift	5
Violation of Ex-Parte Order of Protection	5
Sexual Misconduct - Juvenile	5
Discharge of Firearm (LEO Only)	4
Stealing - Other	4
Unfounded	4
Vehicular - Non-Injury	4
Suicide - Attempted	3
Tampering	3
State Warrant Arrest	3
False Imprisonment	3
Vehicular - Injury	3
Patronizing Prostitution	3
Stealing from Building/Residence	2
Mental Health/Crisis Intervention Team (CIT) Report	2
Attempt to Locate Motor Vehicle	2
Vehicular - Non-Injury Hit and Run	2
Interdiction	2
False Report	2
Service of an Ex-Parte Order of Protection	2
Possession/Sale/Distribution of an Imitation Controlled Substance	1
Suicide	1
Confiscated Firearm	1
Vehicular - Injury Hit and Run	1
Possession of Drug Paraphernalia	1
Unlawful Endangerment of Another	1
Time taken: 11.355 seconds, Fetched: 85 row(s)
hive>
```

Fig 14 Query 8

This query analyzes firearm usage across different offenses by counting incidents where firearms were used. By grouping the data by offense type, it highlights the most prevalent crimes involving firearms, providing essential insights for law enforcement and community safety initiatives.

9. **Top 5 Domestic Violence Flag by offense**

```
select offense,count(*) as DV_offenses
from crime_data
where DVFlag = True
group by offense
order by DV_offenses desc
limit 5;
```

```
hive> select offense,count(*) as DV_offenses
    > from crime_data
    > where DVFlag = True
    > group by offense
    > order by DV_offenses desc
    > limit 5;
Query ID = root_20241102220217_11d2ce4c-b838-461e-997c-7673a7d398ec
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.30 s
----------------------------------------------------------------------------------------------
OK
Domestic Violence Assault (Non-Aggravated)      17434
Domestic Violence Assault (Aggravated)  5736
Domestic Violence Property Damage       1719
Harassment / Intimidation - Domestic Violence   1345
Violation of Full Order of Protection   618
Time taken: 9.77 seconds, Fetched: 5 row(s)
```

Fig 15 Query 9

This query identifies the top five offenses classified as domestic violence by counting incidents marked with the Domestic Violence flag. It provides a focused view of the most common domestic violence offenses, aiding in targeted intervention and prevention efforts within the community.

10. **Offenses by Reported Year and Month (Year-Month Breakdown)**

```
select Reported_year,Reported_month,count(*) as
offenses_by_year_month
from crime_data
group by Reported_year,Reported_month
order by Reported_year,Reported_month;
```

```
hive> select Reported_year,Reported_month,count(*) as offenses_by_year_month
    > from crime_data
    > group by Reported_year,Reported_month
    > order by Reported_year,Reported_month;
Query ID = root_20241102220456_0933edb6-f756-420f-886d-fdadacc1b4ce
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.52 s
----------------------------------------------------------------------------------------------
OK
NULL    NULL    1
2020    1       4000
2020    2       3955
2020    3       3804
2020    4       3447
2020    5       4134
2020    6       4873
```

```
2022    8       6806
2022    9       6697
2022    10      6875
2022    11      5527
2022    12      5484
2023    1       6388
2023    2       5385
2023    3       6318
2023    4       6601
2023    5       7060
2023    6       7479
2023    7       7881
2023    8       7949
2023    9       7408
2023    10      7055
2023    11      6707
2023    12      5910
2024    1       6552
2024    2       6609
2024    3       6764
2024    4       7051
2024    5       7595
2024    6       7782
2024    7       8565
2024    8       8219
2024    9       7341
2024    10      4192
Time taken: 10.978 seconds, Fetched: 59 row(s)
```

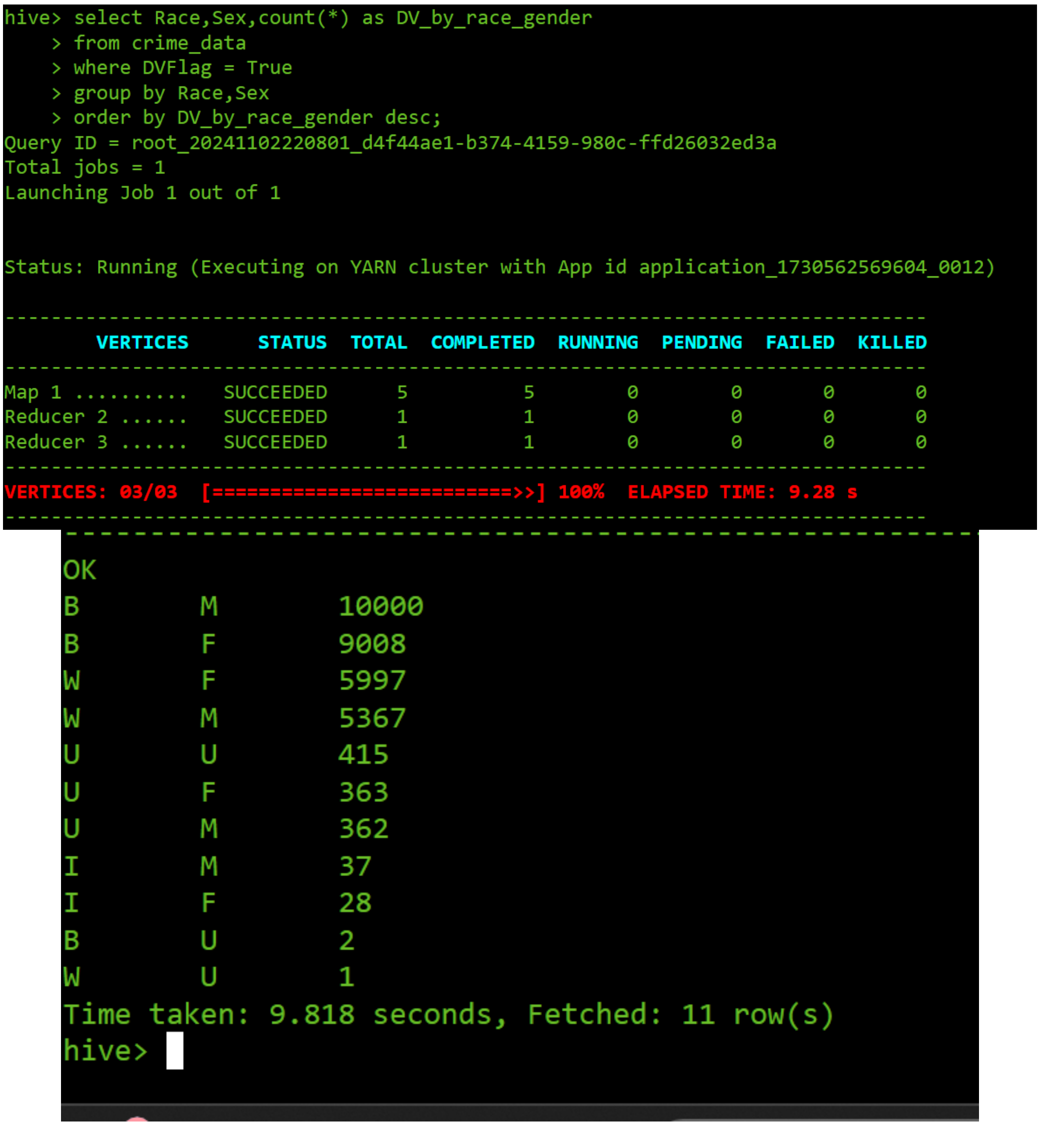
Fig 16 Query 10

This query aggregates the total number of offenses reported each month across different years. By providing a year-month breakdown, it enables the analysis of crime trends over time, helping to identify seasonal patterns and shifts in criminal activity.

11. **Race and Gender Distribution for Domestic Violence Cases**

```
select Race,Sex,count(*) as DV_by_race_gender
from crime_data
```

```
where DVFlag = True
group by Race,Sex
order by DV_by_race_gender desc;
```



```
hive> select Race,Sex,count(*) as DV_by_race_gender
    > from crime_data
    > where DVFlag = True
    > group by Race,Sex
    > order by DV_by_race_gender desc;
Query ID = root_20241102220801_d4f44ae1-b374-4159-980c-ffd26032ed3a
Total jobs = 1
Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......  SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......  SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.28 s
----------------------------------------------------------------------------------------------
OK
B       M       10000
B       F       9008
W       F       5997
W       M       5367
U       U       415
U       F       363
U       M       362
I       M       37
I       F       28
B       U       2
W       U       1
Time taken: 9.818 seconds, Fetched: 11 row(s)
hive>
```

Fig 17 Query 11

This query counts domestic violence incidents categorized by race and gender. By grouping the data in this manner, it reveals patterns of domestic violence across different demographic groups, facilitating targeted interventions and informed policy-making to address these issues.

12. **Count of Unique Offense Descriptions**

```
select count(distinct Description) as unique_desc
from crime_data;
```

```
hive> select count(distinct Description) as unique_desc
    > from crime_data;
Query ID = root_20241102221106_6382fe3c-922b-4803-801f-328ba889b284
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.40 s
----------------------------------------------------------------------------------------------
OK
61
Time taken: 10.045 seconds, Fetched: 1 row(s)
hive>
```

Fig 18 Query 12

This query calculates the number of unique offense descriptions in the crime dataset. By identifying distinct offenses, it provides insights into the variety and complexity of crime types recorded, which can help in understanding crime trends and law enforcement strategies.

13. **Top 5 Most Common Offense Description**

```
select Description,count(*) as desc_count
from crime_data
group by Description
order by desc_count desc
limit 5;
```

```
hive> select Description,count(*) as desc_count
    > from crime_data
    > group by Description
    > order by desc_count desc
    > limit 5;
Query ID = root_20241102221314_c8cbb279-14fb-4370-91c5-9310c7519667
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.52 s
----------------------------------------------------------------------------------------------
OK
Simple Assault  48457
Motor Vehicle Theft     36256
Aggravated Assault      34955
Vandalism/Destruction of Property       34054
Theft From Motor Vehicle        23169
Time taken: 10.026 seconds, Fetched: 5 row(s)
```

Fig 19 Query 13

This query identifies the top five most common offense descriptions in the crime dataset by counting occurrences of each description. It helps highlight prevalent crime types, which can inform public safety initiatives and resource allocation for law enforcement.

14. **Top 5 Most Frequent Reporting District**

```
select Rep_Dist,count(*) as offenses_by_Repdist
from crime_data
group by Rep_Dist
order by offenses_by_Repdist desc
limit 5;
```

```
hive> select Rep_Dist,count(*) as offenses_by_Repdist
    > from crime_data
    > group by Rep_Dist
    > order by offenses_by_Repdist desc
    > limit 5;
Query ID = root_20241102221510_34dff49c-1f72-458d-babb-4b846d562a75
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......  SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......  SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.07 s
----------------------------------------------------------------------------------------------
OK
PJ3601  7788
PP0321  4292
PJ4990  2507
PJ2650  2374
PJ1938  1887
Time taken: 9.522 seconds, Fetched: 5 row(s)
```

Fig 20 Query 14

This query retrieves the five representative districts with the highest number of reported offenses. By analyzing offenses by district, it provides insights into crime concentration areas, aiding law enforcement in targeting resources and strategies effectively.

15. **Top 5 Streets with Most Reported Offenses**

```
select Address,count(*) as offense_by_address
from crime_data
group by Address
order by offense_by_address desc
limit 5;
```

```
hive> select Address,count(*) as offense_by_address
    > from crime_data
    > group by Address
    > order by offense_by_address desc
    > limit 5;
Query ID = root_20241102221807_0e91bb15-415d-4e5e-aba7-1b348b651dd9
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.44 s
----------------------------------------------------------------------------------------------
OK
11600  E US 40 HWY      5658
8500  N BOARDWALK AVE   3531
100  E LINWOOD BLVD     2294
2300  HOLMES ST 1782
2300  E MEYER BLVD      1679
Time taken: 10.963 seconds, Fetched: 5 row(s)
```

Fig 21 Query 15

This query identifies the top five addresses with the highest frequency of reported offenses. By pinpointing high-crime locations, it assists in understanding crime hotspots and enables law enforcement to allocate resources and implement preventive measures more effectively.

16. **Offenses by Reported Day of the Week**

```
select from_unixtime(unix_timestamp(concat(Reported_year, '-', 
Reported_month, '-', Reported_day), 'yyyy-MM-dd'), 'EEEE') AS 
day_of_week,
COUNT(*) AS offenses_by_dayofweek
from crime_data
group by from_unixtime(unix_timestamp(concat(Reported_year, '-', 
Reported_month, '-', Reported_day), 'yyyy-MM-dd'), 'EEEE')
order by offenses_by_dayofweek DESC;
```

```
hive> select from_unixtime(unix_timestamp(concat(Reported_year, '-', Reported_month, '-', Reported_day), 'yyyy-MM-dd'), 'EEEE') AS day_of_we
ek,
    > COUNT(*) AS offenses_by_dayofweek
    > from crime_data
    > group by from_unixtime(unix_timestamp(concat(Reported_year, '-', Reported_month, '-', Reported_day), 'yyyy-MM-dd'), 'EEEE')
    > order by offenses_by_dayofweek DESC;
Query ID = root_20241102222456_d25025aa-3c72-4d6b-934e-c547237f0289
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0012)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.82 s
----------------------------------------------------------------------------------------------
```

```
OK
Monday  53007
Friday  50580
Tuesday 50367
Wednesday       49985
Thursday        48491
Saturday        48482
Sunday  48441
NULL    1
Time taken: 12.441 seconds, Fetched: 8 row(s)
hive>
```

Fig 22 Query 16

This query analyzes the distribution of offenses by day of the week, counting the total offenses for each day. By identifying trends related to specific days, it aids in understanding crime patterns and optimizing policing strategies for higher crime days.

17. **Percentage of Offenses Involving Firearms**

```
SELECT (COUNT(IF(Firearm_Used_Flag = TRUE, 1, NULL)) * 100.0) / 
COUNT(*) AS Firearm_Usage_per FROM crime_data;
```

```
hive> SELECT (COUNT(IF(Firearm_Used_Flag = TRUE, 1, NULL)) * 100.0) / COUNT(*) AS Firearm_Usage_per FROM crime_data;
Query ID = root_20241102223836_60de4d53-0519-46a1-8134-b75237ba86c6
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1730562569604_0013)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.00 s
----------------------------------------------------------------------------------------------
OK
9.044980163387281
Time taken: 17.14 seconds, Fetched: 1 row(s)
hive>
```

Fig 23 Query 17

This query calculates the percentage of offenses where firearms were used by comparing the count of firearm-related offenses to the total number of offenses. It provides insight into the prevalence of firearm use in crimes, which is crucial for assessing public safety and policy implications.

18. **Age Distribution for Top 3 Offenses**

```
SELECT Offense, Age, COUNT(*) AS Age_Distribution
from crime_data
where offense in ('Stolen Auto','Domestic Violence Assault (Non-Aggravated)','Property Damage')
group by offense,Age
order by offense,Age;
```

```
hive> SELECT Offense, Age, COUNT(*) AS Age_Distribution
    > from crime_data
    > where offense in ('Stolen Auto','Domestic Violence Assault (Non-Aggravated)','Property Damage')
    > group by offense,Age
    > order by offense,Age;
Query ID = root_20241102225051_6f597b5c-a886-46e6-82ec-6503593c1d91
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0013)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.96 s
----------------------------------------------------------------------------------------------
OK
Domestic Violence Assault (Non-Aggravated)      NULL    94
Domestic Violence Assault (Non-Aggravated)      18.0    486
Domestic Violence Assault (Non-Aggravated)      19.0    596
Domestic Violence Assault (Non-Aggravated)      20.0    723
Domestic Violence Assault (Non-Aggravated)      21.0    777
Domestic Violence Assault (Non-Aggravated)      22.0    858
```

```
Domestic Violence Assault (Non-Aggravated)      83.0    4
Domestic Violence Assault (Non-Aggravated)      84.0    2
Domestic Violence Assault (Non-Aggravated)      85.0    7
Domestic Violence Assault (Non-Aggravated)      86.0    5
Domestic Violence Assault (Non-Aggravated)      87.0    5
Domestic Violence Assault (Non-Aggravated)      88.0    2
Domestic Violence Assault (Non-Aggravated)      89.0    1
Domestic Violence Assault (Non-Aggravated)      91.0    1
Domestic Violence Assault (Non-Aggravated)      92.0    4
Domestic Violence Assault (Non-Aggravated)      94.0    1
Domestic Violence Assault (Non-Aggravated)      97.0    1
Property Damage NULL    4
Property Damage 18.0    152
Property Damage 19.0    211
Property Damage 20.0    315
Property Damage 21.0    406
Property Damage 22.0    437
Property Damage 23.0    555
Property Damage 24.0    507
Property Damage 25.0    618
Property Damage 26.0    569
Property Damage 27.0    552
Property Damage 28.0    562
Property Damage 29.0    543
Property Damage 30.0    681
Property Damage 31.0    565
Property Damage 32.0    487
Property Damage 33.0    544
Property Damage 34.0    475
```

Fig 24 Query 18

```
Property Damage 77.0    38
Property Damage 78.0    37
Property Damage 79.0    41
Property Damage 80.0    38
Property Damage 81.0    27
Property Damage 82.0    16
Property Damage 83.0    15
Property Damage 84.0    23
Property Damage 85.0    9
Property Damage 86.0    14
Property Damage 87.0    7
Property Damage 88.0    8
Property Damage 89.0    8
Property Damage 90.0    4
Property Damage 91.0    5
Property Damage 93.0    3
Property Damage 94.0    2
Property Damage 95.0    1
Property Damage 96.0    1
Property Damage 98.0    1
Property Damage 99.0    5
Stolen Auto     NULL    3
Stolen Auto     18.0    303
Stolen Auto     19.0    385
Stolen Auto     20.0    508
Stolen Auto     21.0    623
Stolen Auto     22.0    710
Stolen Auto     23.0    810
Stolen Auto     24.0    850
```

```
Stolen Auto     73.0    105
Stolen Auto     74.0    90
Stolen Auto     75.0    107
Stolen Auto     76.0    61
Stolen Auto     77.0    72
Stolen Auto     78.0    68
Stolen Auto     79.0    61
Stolen Auto     80.0    36
Stolen Auto     81.0    50
Stolen Auto     82.0    28
Stolen Auto     83.0    33
Stolen Auto     84.0    23
Stolen Auto     85.0    25
Stolen Auto     86.0    14
Stolen Auto     87.0    12
Stolen Auto     88.0    11
Stolen Auto     89.0    5
Stolen Auto     90.0    4
Stolen Auto     91.0    3
Stolen Auto     92.0    1
Stolen Auto     93.0    1
Stolen Auto     94.0    3
Stolen Auto     95.0    1
Stolen Auto     97.0    2
Stolen Auto     98.0    2
Stolen Auto     99.0    2
Time taken: 11.44 seconds, Fetched: 240 row(s)
hive>
```

This query analyses the age distribution of offenders for specific offenses: 'Stolen Auto,' 'Domestic Violence Assault (Non-Aggravated),' and 'Property Damage.' It groups the data by offense and age, providing a detailed view of how age correlates with these particular crimes, which can aid in targeted prevention strategies.

19. **Get the Top 5 Areas with the Most Offenses Per Month**

```
SELECT Reported_month, Area, offenses_bymonth_area
from (
SELECT Reported_month, Area, COUNT(*) AS
offenses_bymonth_area,ROW_NUMBER() OVER (PARTITION BY
Reported_month ORDER BY COUNT(*) DESC)
AS rank
```

```
from crime_data
GROUP BY Reported_month, Area
) ranked_data
where rank <=5
ORDER BY Reported_month, offenses_bymonth_area DESC;
```

```
hive> SELECT Reported_month, Area, offenses_bymonth_area
    > from (
    > SELECT Reported_month, Area, COUNT(*) AS offenses_bymonth_area,ROW_NUMBER() OVER (PARTITION BY Reported_month ORDER BY COUNT(*) DESC)
AS rank
    > from crime_data
    > GROUP BY Reported_month, Area
    > ) ranked_data
    > where rank <=5
    > ORDER BY Reported_month, offenses_bymonth_area DESC;
Query ID = root_20241102230025_d1e5721a-665c-4b76-9cc1-d5f2827aee20
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1730562569604_0013)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 4 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 10.70 s
----------------------------------------------------------------------------------------------
```

```
1       CPD     7836
1       EPD     7073
1       MPD     5300
1       SPD     3013
1       NPD     2415
2       CPD     7271
2       EPD     6542
2       MPD     4891
2       SPD     2706
2       NPD     2113
3       CPD     7803
3       EPD     6886
3       MPD     5511
3       SPD     3135
3       NPD     2371
4       CPD     7592
4       EPD     7511
4       MPD     5681
4       SPD     3121
4       NPD     2340
5       CPD     8817
5       EPD     7851
5       MPD     6120
5       SPD     3271
5       NPD     2676
6       CPD     9061
6       EPD     8432
6       MPD     6678
6       SPD     2956
```

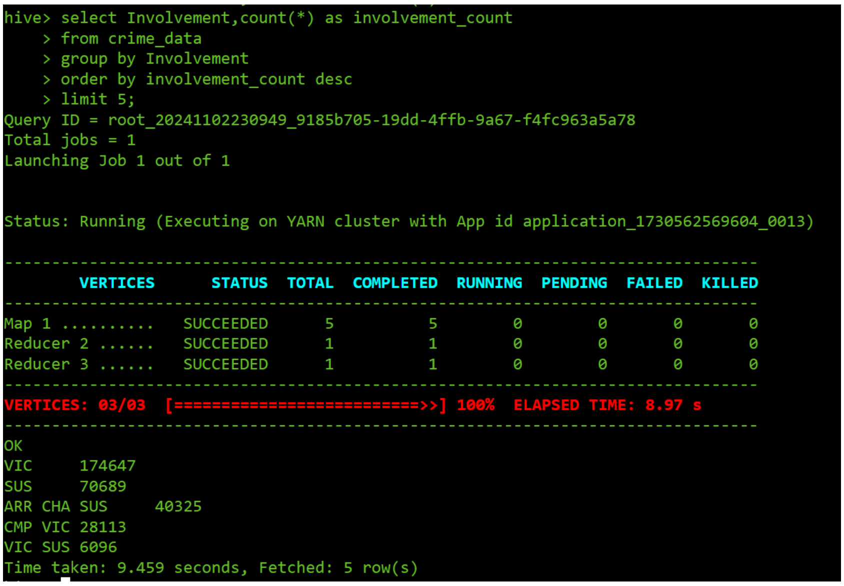
Fig 25 Query 19

```
7       MPD     6586
7       SPD     3686
7       NPD     3028
8       CPD     9805
8       EPD     8563
8       MPD     7005
8       SPD     3495
8       NPD     3022
9       CPD     9323
9       EPD     8238
9       MPD     6457
9       SPD     3420
9       NPD     2943
10      CPD     8258
10      EPD     7261
10      MPD     5758
10      SPD     3101
10      NPD     2508
11      CPD     6245
11      EPD     5800
11      MPD     4461
11      SPD     2370
11      NPD     1885
12      CPD     5519
12      EPD     4739
12      MPD     3909
12      SPD     2244
12      NPD     1773
Time taken: 11.275 seconds, Fetched: 61 row(s)
```

This query extracts the top five areas with the highest offense counts for each month. By ranking the areas based on their offense totals, it provides insights into monthly crime patterns, allowing law enforcement and policymakers to allocate resources effectively and address community concerns regarding safety.

20. **Top 5 Most Common Involvement Type**

```
select Involvement,count(*) as involvement_count
from crime_data
group by Involvement
order by involvement_count desc
limit 5;
```



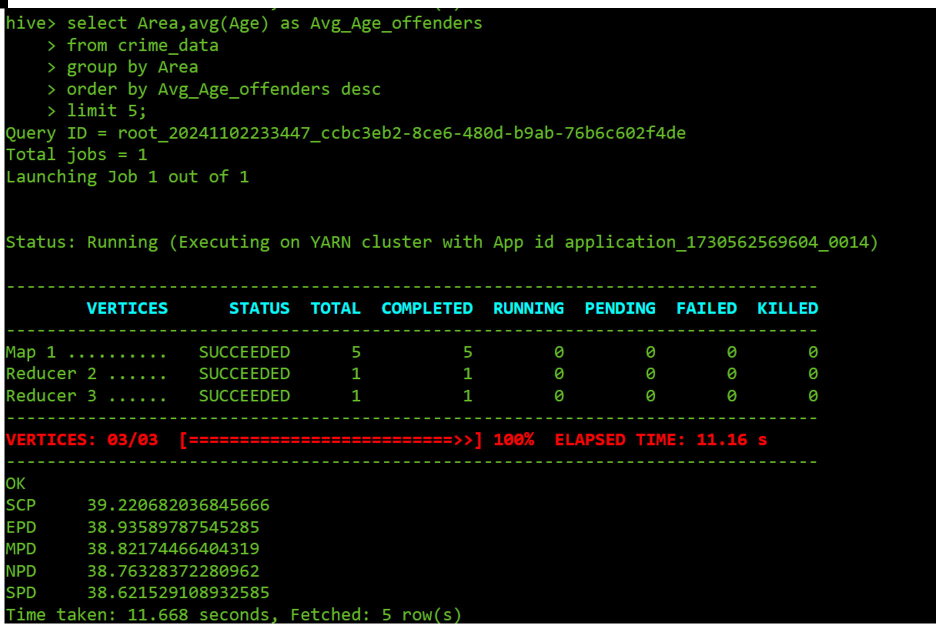
```
hive> select Involvement,count(*) as involvement_count
    > from crime_data
    > group by Involvement
    > order by involvement_count desc
    > limit 5;
Query ID = root_20241102230949_9185b705-19dd-4ffb-9a67-f4fc963a5a78
Total jobs = 1
Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1730562569604_0013)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 8.97 s
----------------------------------------------------------------------------------
OK
VIC     174647
SUS     70689
ARR CHA SUS     40325
CMP VIC 28113
VIC SUS 6096
Time taken: 9.459 seconds, Fetched: 5 row(s)
```

Fig 26 Query 20

This query identifies the five most common types of involvement in crimes by counting occurrences in the dataset. By grouping and sorting the data based on involvement, it highlights prevalent roles—such as victim, suspect, or witness—within criminal incidents, aiding in understanding crime dynamics and community impact.

21. Top 5 Areas with Highest Average Offender Age

```
select Area,avg(Age) as Avg_Age_offenders
from crime_data
group by Area
order by Avg_Age_offenders desc
limit 5;
```



```
hive> select Area,avg(Age) as Avg_Age_offenders
    > from crime_data
    > group by Area
    > order by Avg_Age_offenders desc
    > limit 5;
Query ID = root_20241102233447_ccbc3eb2-8ce6-480d-b9ab-76b6c602f4de
Total jobs = 1
Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1730562569604_0014)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.16 s
----------------------------------------------------------------------------------
OK
SCP     39.220682036845666
EPD     38.93589787545285
MPD     38.82174466404319
NPD     38.76328372280962
SPD     38.621529108932585
Time taken: 11.668 seconds, Fetched: 5 row(s)
```

Fig 27 Query 21

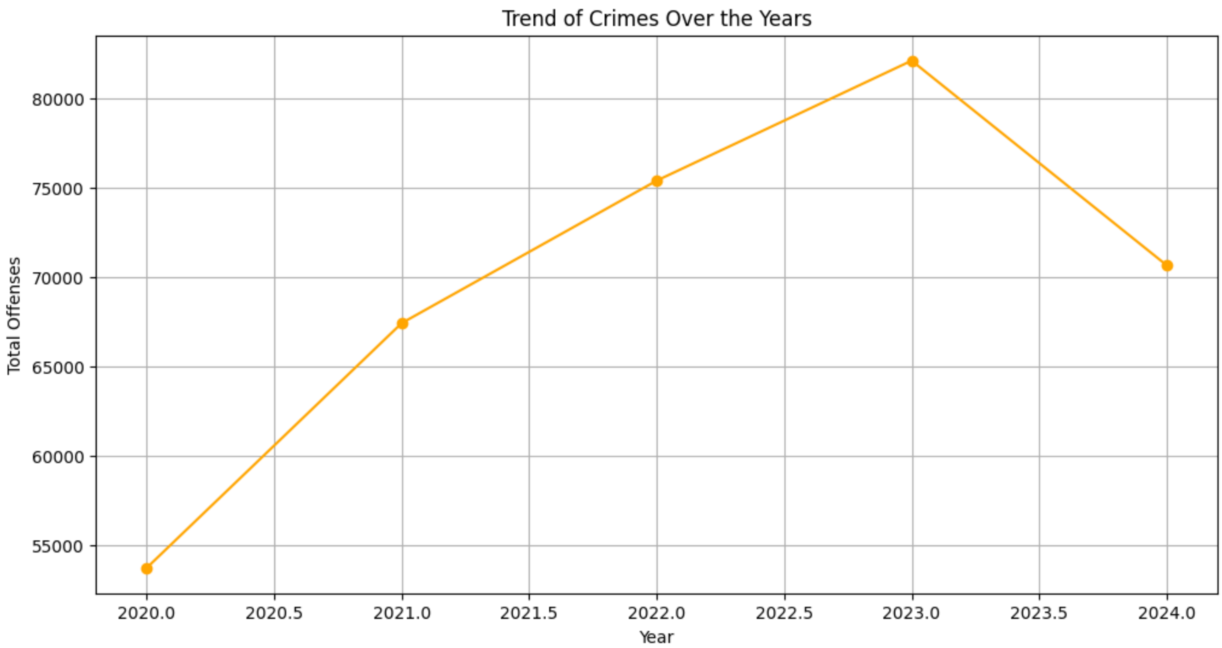
This query calculates the average age of offenders in different areas, providing insight into demographic trends related to crime. By grouping the data by area and sorting by average age, it identifies the top five areas with the oldest offenders, which can inform targeted interventions and resource allocation in law enforcement and community programs.

3.4 Visualization of the Data: -

```
[1]
from pyspark.sql import SparkSession
# Create a Spark session
spark = SparkSession.builder.appName("KCCrime").getOrCreate()
```

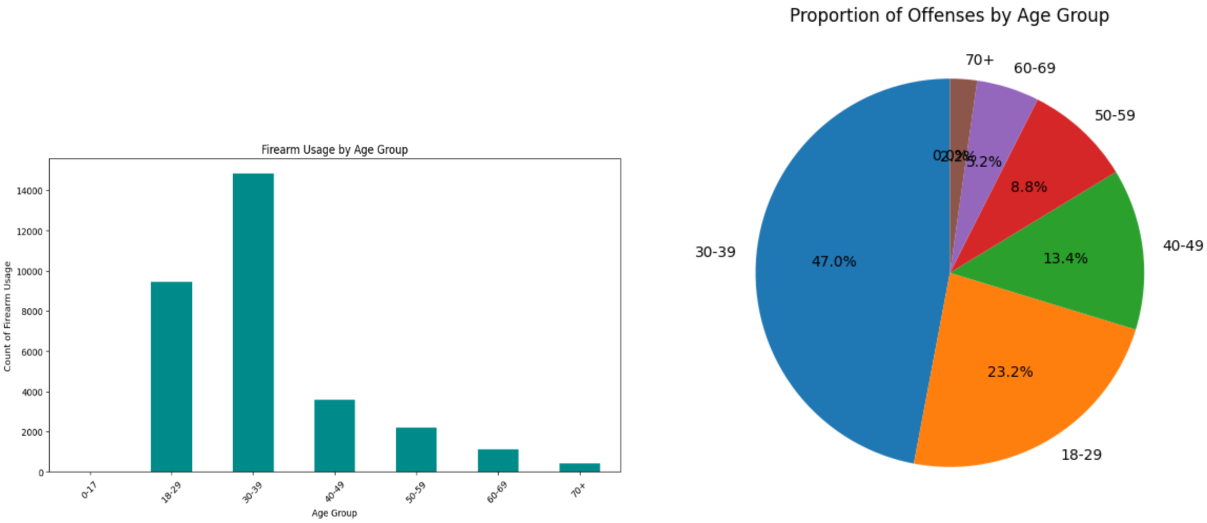
visual 1 pyspark initialization

Initiating the pyspark session to do the visualization of the Crime data.



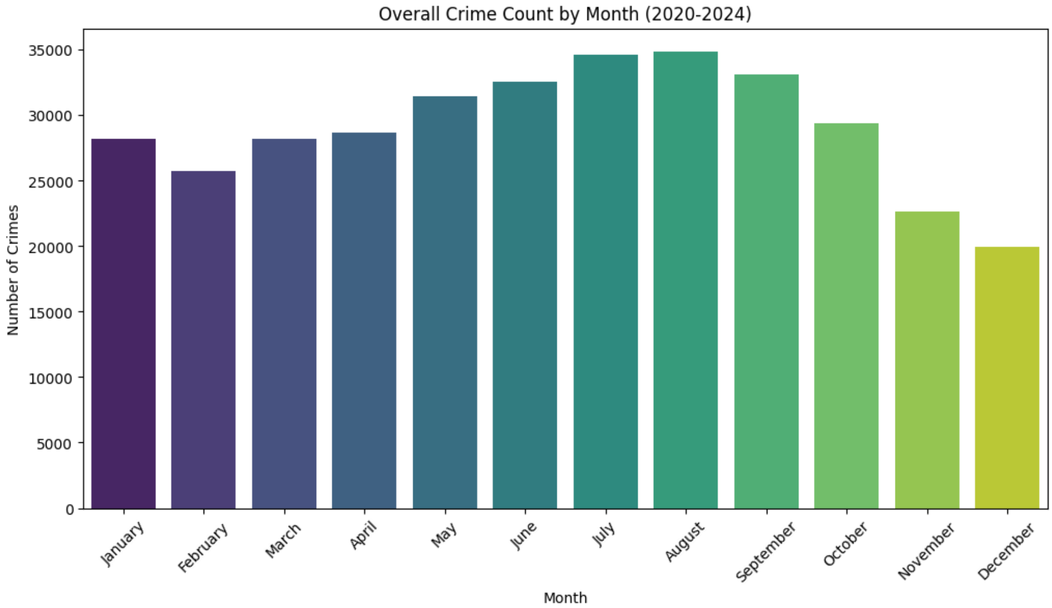
visual 2 crime over years

The Line plot above represents the Year wise crimes and we can observe that during the year 2023 the crime rate was very high. Over the years of 2020 to 2023 there has been a continuous increase in the crimes and the total year data of 2024 is not collected but till October 2024 we can observe that the number of offenses is less



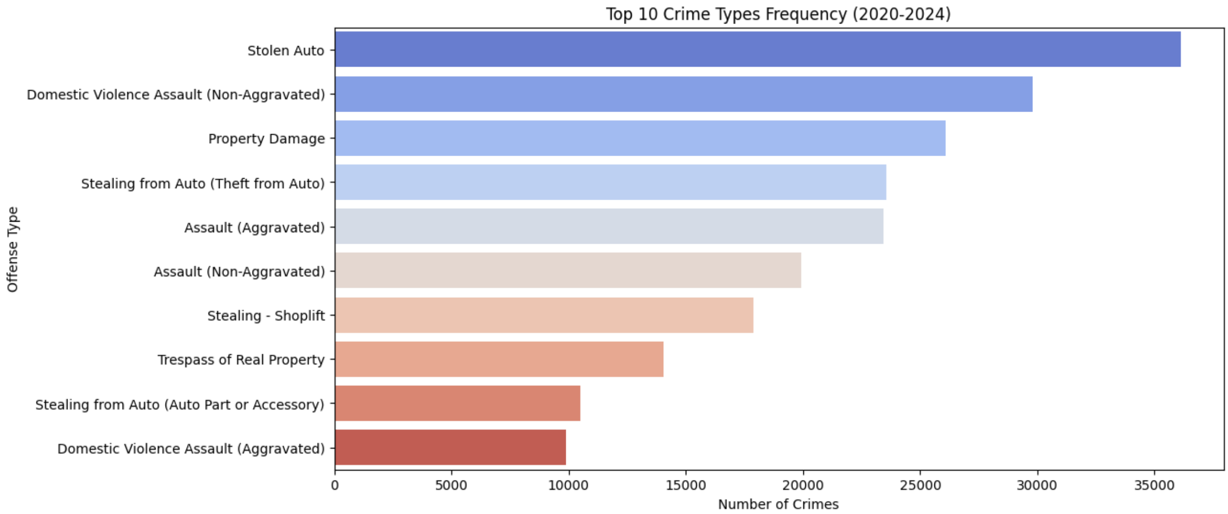
visual 3 Crimes over the age groups

We observe that the age group of 30-39 has access to a Firearms and they do crimes with them. So we need to make sure to check all the people in that age group so that the crimes related to the firearm's reduces.



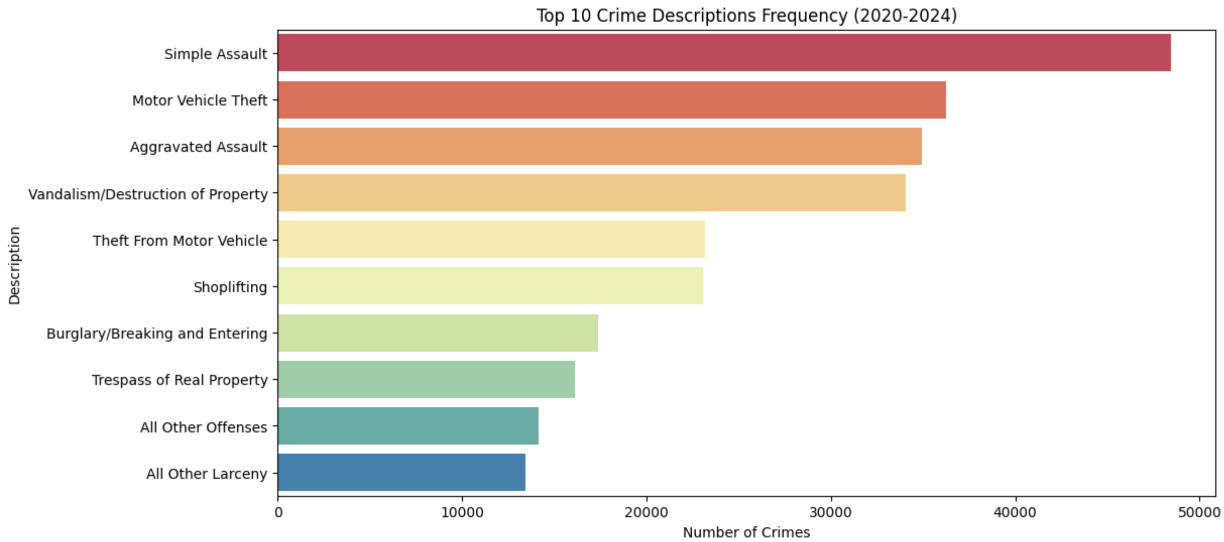
visual 4 Crimes over the month

The above image represents that there is a high crime in august and during the period of June – September the crime rate has been high, least during the year December. We can draw a conclusion that the police should be on high alert during these months.



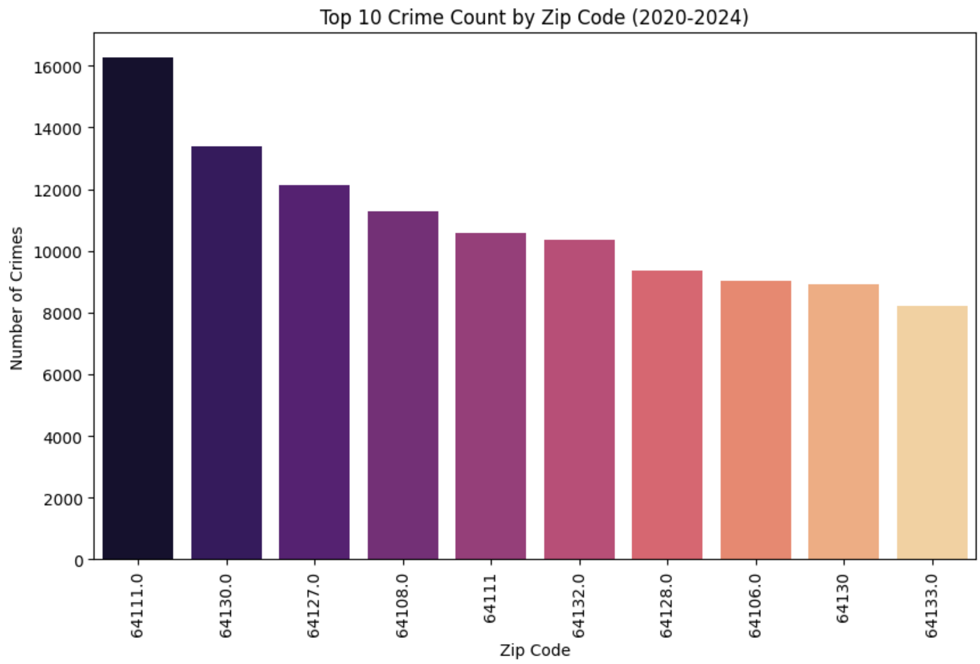
visual 5 Crimes over the type of offenses

Stolen Auto's Crime is the highest type of offense that's occurred over the years.



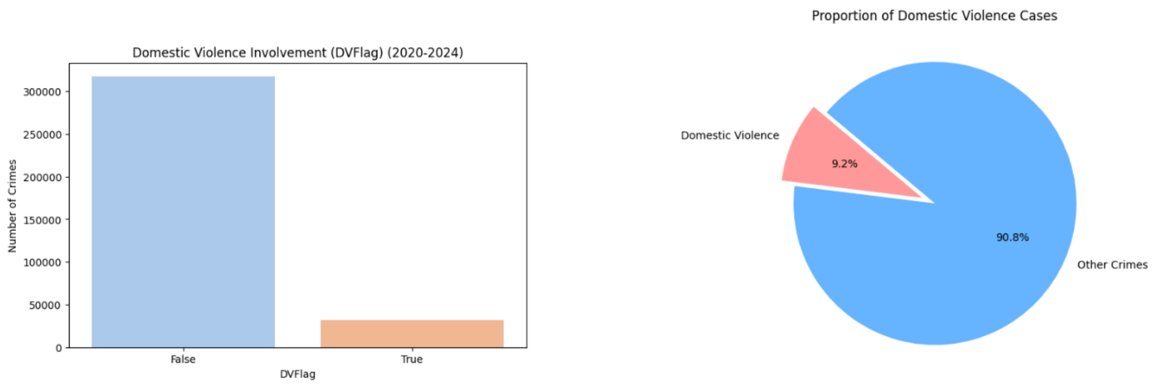
visual 6 Crimes over the descriptions

Here we can see that the word simple Assault is the most common and the highest number of crime that has been occurring.



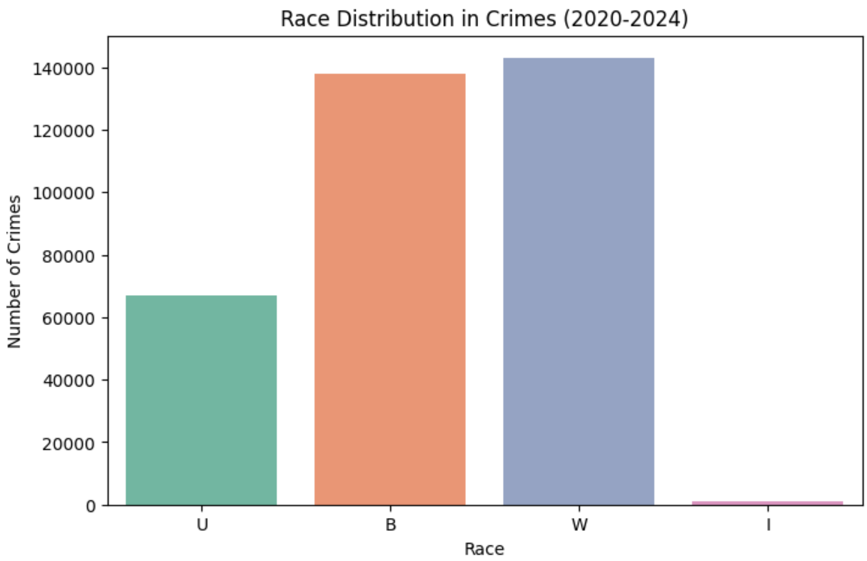
visual 7 Crimes over Zip-code

In the above graph we observe that the number of crimes is high in the zip-codes 64111 and 64130 so there is a need to increase the patrols, the number of cops in that area and surveillance should also increase to reduce the number of crimes.



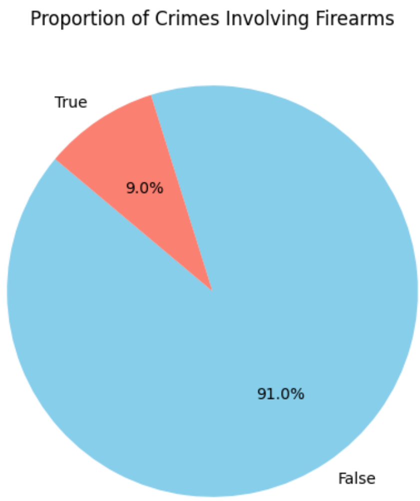
visual 8 Domestic violence crimes

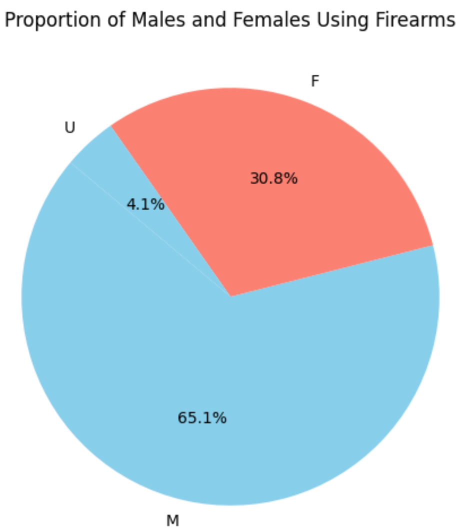
We can clearly observe the difference in the crimes that represents that there are very few domestic crimes that are happening but stull over 40k crimes relating to the domestic violence have happened over the period of time. Its 9.2% is the no of domestic crimes that are happening.



visual 9 Crime - Race

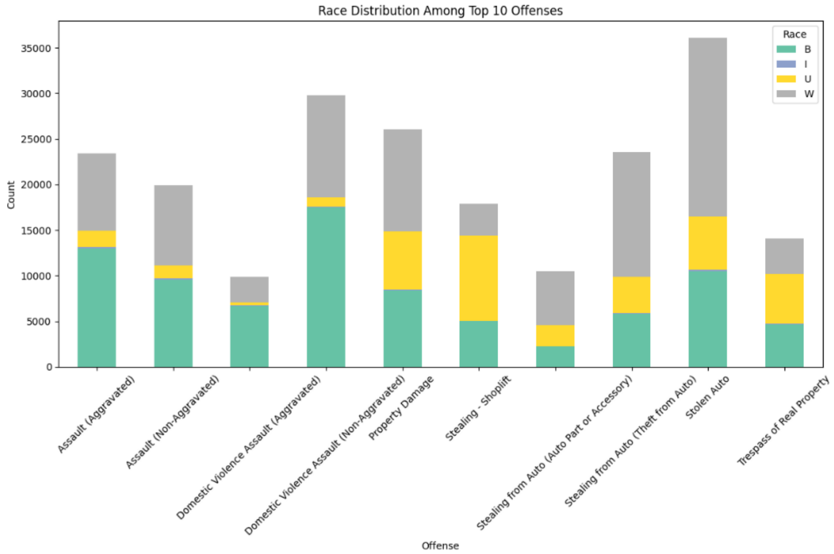
So we observe that we have mainly two races and there are 70k crimes where the race of the person who's done the crime is not mentioned. This represents that we lack in data.





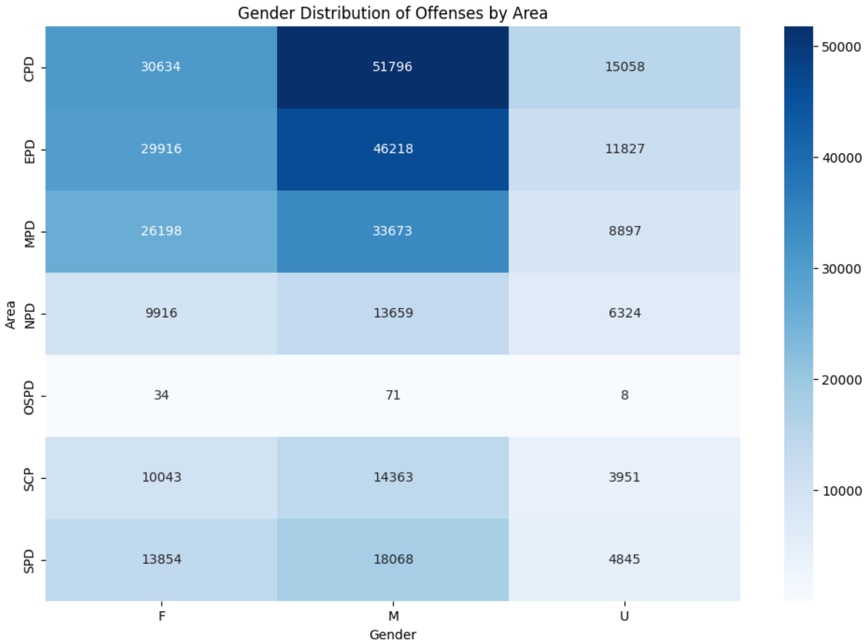
visual 10 Percentage of crime involving Firearms

There is 9% of crimes that are happening while using the firearms. And the no of females and males in that 9% is represented to the right.



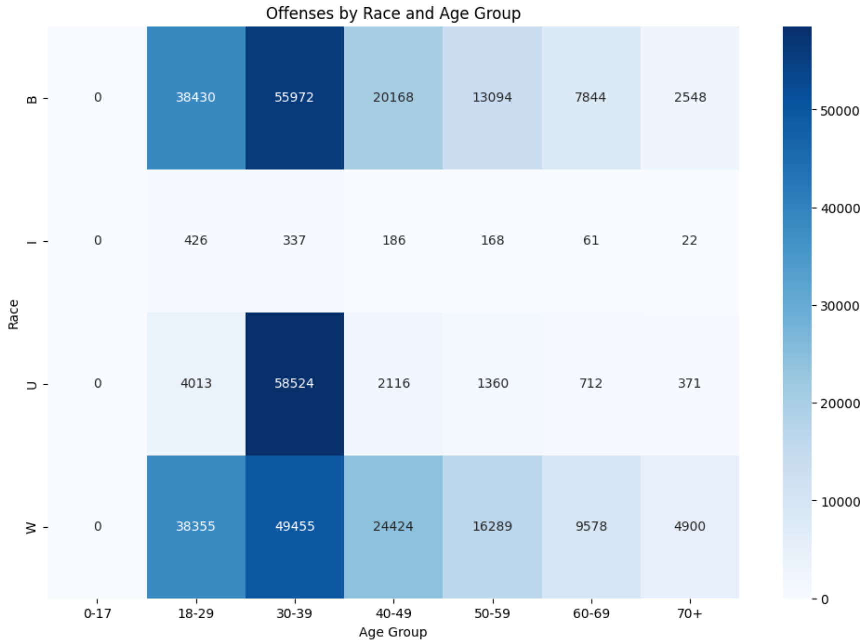
visual 11 Race Distribution among Top 10 crimes.

The graph represents the Crimes and their respective ratio of races that have caused that respective crime. We see that the main race that's causing the stolen auto crime is the White race.



visual 12 Gender Distribution over area

The heat-map represents that there is high number of crime-offenses in the area 'CPD', which implicates that the cops should try to do better at reducing the crimes and increase the surveillance. The Area OSPD is very friendly environment having the lease no of crimes.



visual 13 Offenses by race and age-group

As we have already seen that the crimes in the age group was already high in the age group 30-39 we can observe that there are all the races that have done the crimes but the crimes done by the boys is a bit higher than that of the unknown and the women.

4. Results

In this project, we utilized Apache Hive, a data warehousing tool built on Hadoop, to conduct a comprehensive analysis of crime data. Hive is highly suitable for processing large datasets due to its use of MapReduce programming, allowing us to efficiently handle vast volumes of data. By leveraging Hive's SQL-like querying language, we could retrieve meaningful insights that would be otherwise challenging to extract manually or through traditional databases. This enabled us to identify patterns, correlations, and trends within the dataset, helping to bring clarity to complex crime data in an accessible and organized manner.

Our findings provide insights that can inform targeted interventions, such as implementing youth awareness programs and enhancing community safety initiatives. Observations about trends by age group, offense type, and temporal factors indicate specific areas where proactive measures may mitigate crime. For instance, identifying peak hours of offenses and demographic-specific trends highlights key risk periods and populations that could benefit from preventive actions, like educational outreach or increased patrolling during specific times and in specific areas. Such data-driven decision-making can improve resource allocation, making crime prevention efforts both more efficient and effective.

Finally, for data visualization, we used Python, a versatile tool for graphical representation that allowed us to transform raw, complex data into clear, visually engaging charts and graphs. This approach made our findings more accessible and easier to interpret compared to traditional tabular formats, providing a more intuitive understanding of crime trends and risk factors. Visualizations like bar charts, line graphs, and heatmaps made it possible to spot patterns briefly, fostering a better grasp of the data's story. This combination of Hive's processing power and Python's visualization capabilities allowed us to convert large-scale data into actionable insights, providing valuable contributions to research on crime prevention and community safety

5. Concluding Remarks

This report provides an analysis of crime data utilizing Hive for data processing and PySpark for visualization, emphasizing the extraction of insights related to crime trends and demographics. The primary objective was to leverage Hive's capabilities to efficiently analyze large datasets while employing PySpark to create visually impactful representations of the findings.

The analysis produced several key insights:

Data Insights: The investigation revealed critical trends in crime occurrences, revealing how various demographic factors and types of offenses correlate with crime rates. Specific demographics were found to be associated with higher instances of certain crimes, such as domestic violence and firearm usage, highlighting the need for targeted community responses.

Visualization Impact: By employing PySpark for visualization, complex datasets were transformed into clear, engaging graphics that enhance understanding. Visualizations such as bar charts, heatmaps, and trend lines effectively communicated the underlying narratives, making it easier for stakeholders to grasp essential patterns and relationships.

Practical Implications: The insights gleaned from this analysis can inform targeted interventions aimed at crime reduction, such as awareness campaigns for at-risk groups and enhanced security measures in identified hotspots. The integration of efficient data processing through Hive with compelling visualizations in PySpark not only deepens our understanding of crime dynamics but also supports the formulation of proactive strategies for crime prevention and community engagement.

6. Future Work

For future work, there are several potential avenues to explore:

1. **Predictive Modelling**: Develop machine learning models to forecast crime occurrences using historical data. By leveraging variables like time, location, and demographics, these models can enhance resource allocation for law enforcement and inform proactive measures.
2. **Geospatial Analysis**: Utilize Geographic Information Systems (GIS) to visualize crime hotspots and spatial patterns. This analysis helps identify high-crime areas, enabling targeted interventions and community policing efforts.
3. **Real-Time Data Integration**: Create a system to incorporate real-time crime data from various sources, such as police reports and social media. This approach allows for timely analysis and improves situational awareness for law enforcement.
4. **Socioeconomic Data Enrichment**: Integrate external datasets like socioeconomic indicators to analyse correlations with crime rates. This can provide insights into the underlying causes of crime and guide community development strategies.
5. **Crime Severity Index Development**: Establish an index to assess crime severity, combining factors like offense nature and victim impact. This helps law enforcement prioritize cases and allocate resources effectively.
6. **Temporal Analysis of Crime Patterns**: Investigate crime trends based on time of day and seasonality. Identifying peak crime times can inform patrol strategies and community engagement efforts.
7. **Community Outreach Programs**: Design community outreach initiatives focused on crime prevention based on data analysis findings. These programs can empower residents and foster collaboration with law enforcement.
8. **Collaboration with Urban Planning**: Partner with urban planners to explore the relationship between neighbourhood design and crime rates. Insights can guide urban development strategies to reduce crime opportunities.
9. **Longitudinal Studies on Crime Trends**: Conduct studies to observe how crime patterns evolve over time. This research can evaluate intervention effectiveness and provide insights for future prevention strategies.

10. **Public Policy Recommendations**: Develop recommendations for policymakers based on data analysis findings. This could involve assessing legislation impacts on crime rates and advocating for community-based safety approaches.

7. References

1. https://data.kcmo.org/browse?category=Crime
2. https://colab.research.google.com/drive/1y19B8Rlv3pzSjx2QpjExyC084hH7r74W?usp=sharing
3. https://cwiki.apache.org/confluence/display/Hive/