Los Angeles and New York crime dataset analysis

Course: BIG DATA TOOLS & TECHNO. II
Prof. Nizar Ali

Group 12 members:

Jose Anicio Pereira Junior - ID: 101506160
Himanshu Kakkar - ID: 101510716
Pujan Bhatt - ID: 101527283

Table of contents

- Project overview
- Objectives
- Dataset and sources
- Data flow chart
- Data Preprocessing and transformation
- Results and graphs
- Conclusion

Project overview

This project focuses on analyzing and comparing crime rates in Los Angeles and New York City using extensive datasets provided by the respective police departments.

The aim is to integrate and process these datasets using various big data tools and technologies to uncover patterns and insights into crime dynamics in these two major cities.

Objectives

The primary objective of this project is to compare crime rates between Los Angeles and New York City by analyzing large datasets.

This analysis aims to identify patterns, trends, and insights into the crime dynamics in these two major cities.

Dataset and sources

LYPD DATA

Arrest Data from 2010 to 2019 -: https://catalog.data.gov/dataset/arrest-data-from-2010-to-2019

Crime Data from 2020 to Present -:https://catalog.data.gov/dataset/crime-data-from-2020-to-present

LAPD Calls for Service 2019-: https://catalog.data.gov/dataset/lapd-calls-for-service-2019

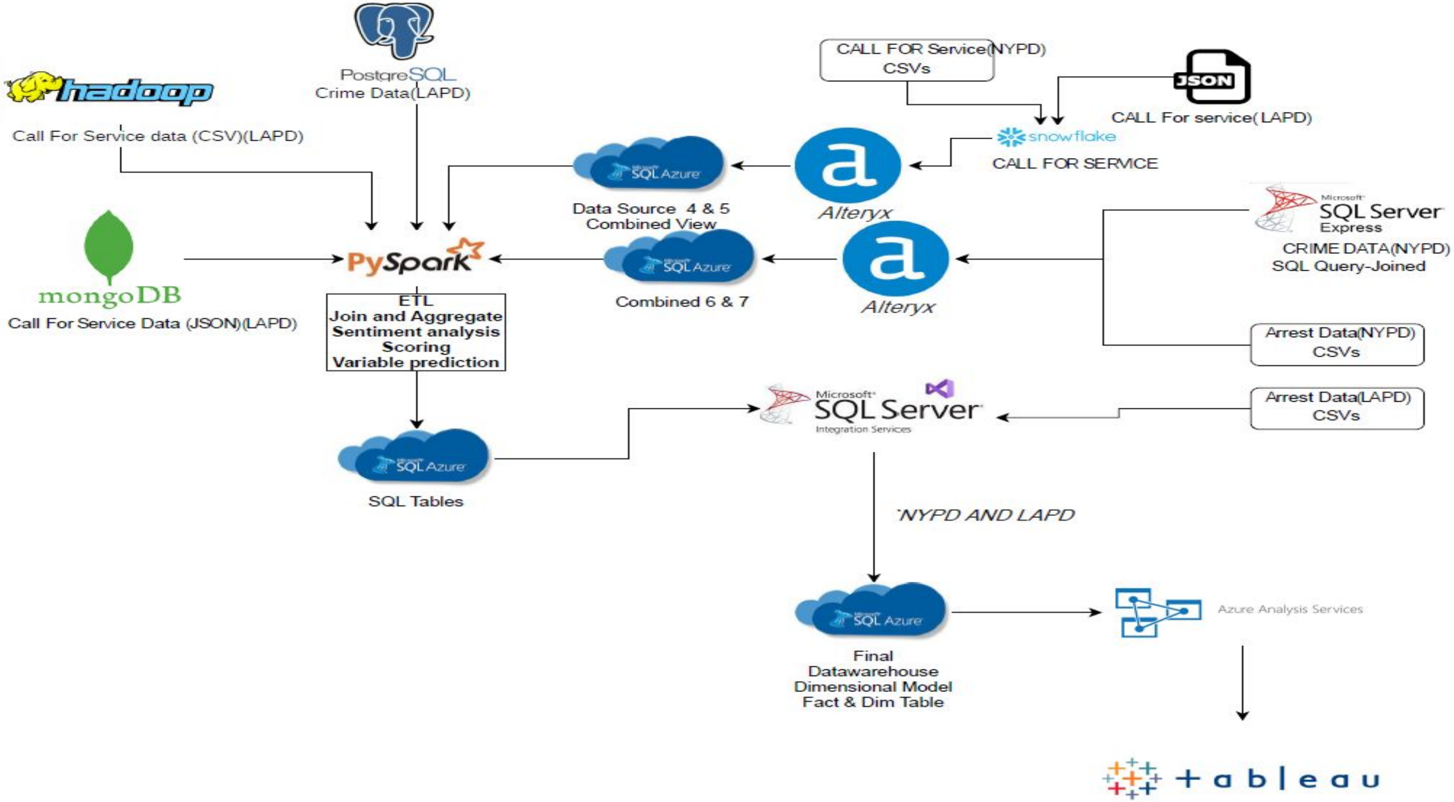
NEW YORK DATA

Arrest Data -: https://catalog.data.gov/dataset/nypd-arrests-data-historic

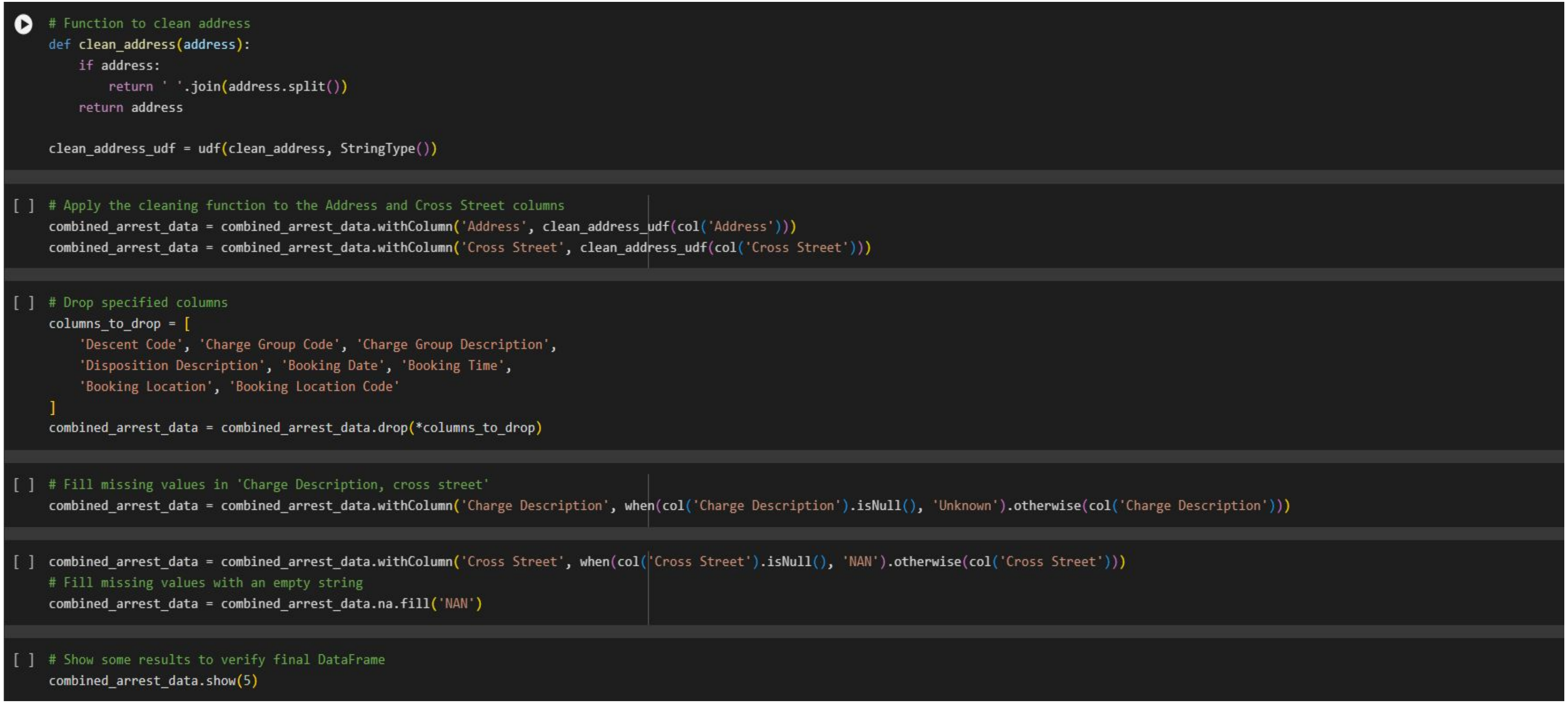
Call for Service Data -: https://catalog.data.gov/dataset/nypd-calls-for-service-historic

Crime Data -: https://catalog.data.gov/dataset/nypd-complaint-data-historic

Data flow chart



Data Preprocessing and transformation



```
# Function to clean address
def clean_address(address):
    if address:
        return ' '.join(address.split())
    return address

clean_address_udf = udf(clean_address, StringType())
```

```
# Apply the cleaning function to the Address and Cross Street columns
combined_arrest_data = combined_arrest_data.withColumn('Address', clean_address_udf(col('Address')))
combined_arrest_data = combined_arrest_data.withColumn('Cross Street', clean_address_udf(col('Cross Street')))
```

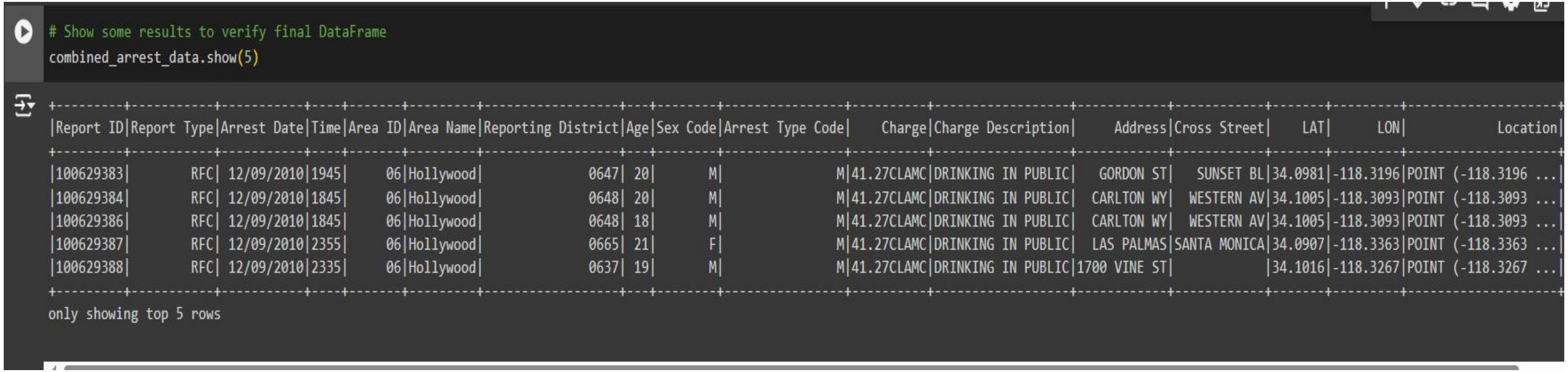
```
# Drop specified columns
columns_to_drop = [
    'Descent Code', 'Charge Group Code', 'Charge Group Description',
    'Disposition Description', 'Booking Date', 'Booking Time',
    'Booking Location', 'Booking Location Code'
]
combined_arrest_data = combined_arrest_data.drop(*columns_to_drop)
```

```
# Fill missing values in 'Charge Description, cross street'
combined_arrest_data = combined_arrest_data.withColumn('Charge Description', when(col('Charge Description').isNull(), 'Unknown').otherwise(col('Charge Description')))
```

```
combined_arrest_data = combined_arrest_data.withColumn('Cross Street', when(col('Cross Street').isNull(), 'NAN').otherwise(col('Cross Street')))
# Fill missing values with an empty string
combined_arrest_data = combined_arrest_data.na.fill('NAN')
```

```
# Show some results to verify final DataFrame
combined_arrest_data.show(5)
```

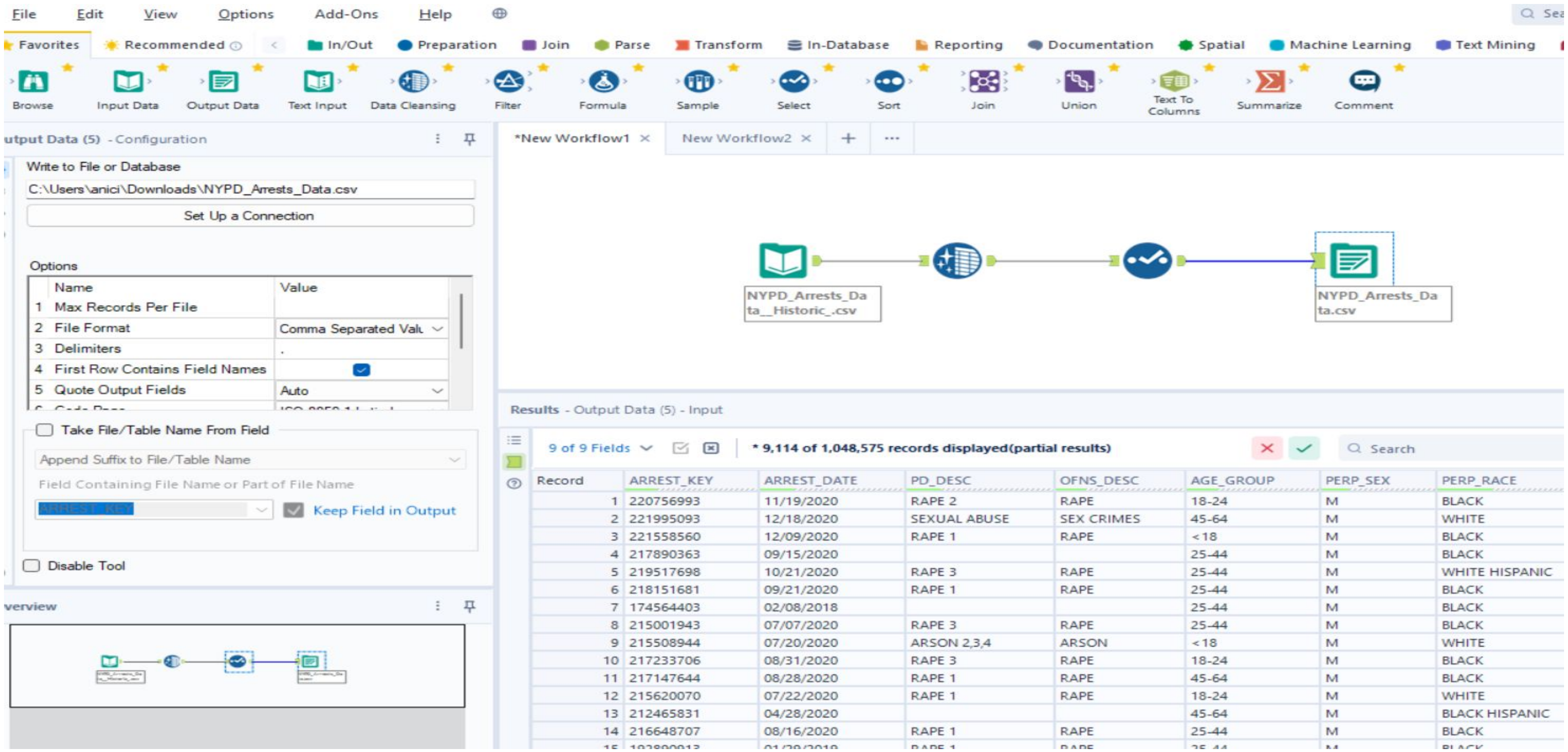
Data Preprocessing and transformation



```
# Show some results to verify final DataFrame
combined_arrest_data.show(5)
```

```
+---------+-----------+-----------+----+-------+---------+------------------+---+--------+----------------+----------+------------------+------------+------------+-------+---------+--------------------+
|Report ID|Report Type|Arrest Date|Time|Area ID|Area Name|Reporting District|Age|Sex Code|Arrest Type Code|    Charge|Charge Description|     Address|Cross Street|    LAT|      LON|            Location|
+---------+-----------+-----------+----+-------+---------+------------------+---+--------+----------------+----------+------------------+------------+------------+-------+---------+--------------------+
|100629383|        RFC| 12/09/2010|1945|     06|Hollywood|              0647| 20|       M|               M|41.27CLAMC|DRINKING IN PUBLIC|   GORDON ST|   SUNSET BL|34.0981|-118.3196|POINT (-118.3196 ...|
|100629384|        RFC| 12/09/2010|1845|     06|Hollywood|              0648| 20|       M|               M|41.27CLAMC|DRINKING IN PUBLIC|  CARLTON WY|  WESTERN AV|34.1005|-118.3093|POINT (-118.3093 ...|
|100629386|        RFC| 12/09/2010|1845|     06|Hollywood|              0648| 18|       M|               M|41.27CLAMC|DRINKING IN PUBLIC|  CARLTON WY|  WESTERN AV|34.1005|-118.3093|POINT (-118.3093 ...|
|100629387|        RFC| 12/09/2010|2355|     06|Hollywood|              0665| 21|       F|               M|41.27CLAMC|DRINKING IN PUBLIC|  LAS PALMAS|SANTA MONICA|34.0907|-118.3363|POINT (-118.3363 ...|
|100629388|        RFC| 12/09/2010|2335|     06|Hollywood|              0637| 19|       M|               M|41.27CLAMC|DRINKING IN PUBLIC|1700 VINE ST|            |34.1016|-118.3267|POINT (-118.3267 ...|
+---------+-----------+-----------+----+-------+---------+------------------+---+--------+----------------+----------+------------------+------------+------------+-------+---------+--------------------+
only showing top 5 rows
```

Data Preprocessing and transformation



utput Data (5) - Configuration

Write to File or Database

C:\Users\anici\Downloads\NYPD_Arrests_Data.csv

Set Up a Connection

Options

	Name	Value
1	Max Records Per File	
2	File Format	Comma Separated Valu
3	Delimiters	,
4	First Row Contains Field Names	☑
5	Quote Output Fields	Auto

☐ Take File/Table Name From Field

Append Suffix to File/Table Name

Field Containing File Name or Part of File Name

ARREST_KEY ☑ Keep Field in Output

☐ Disable Tool

Results - Output Data (5) - Input

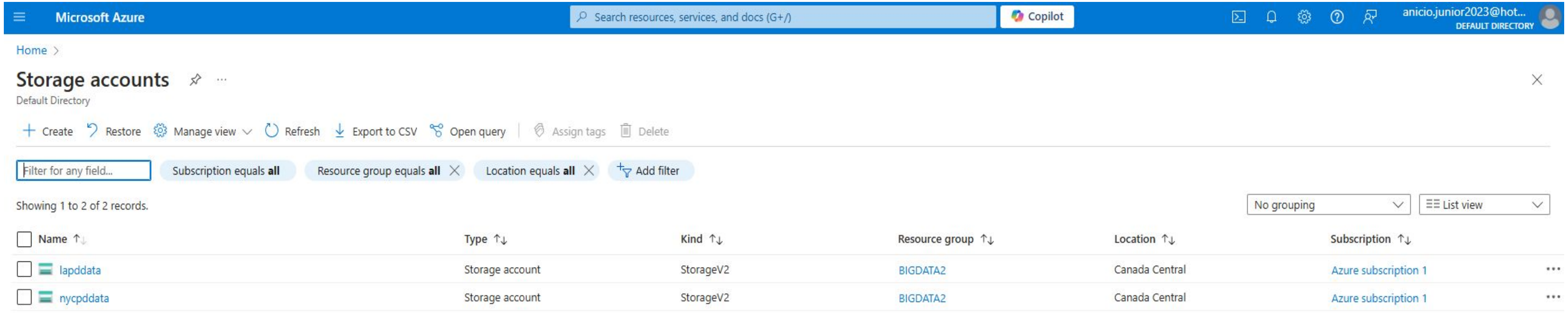
9 of 9 Fields

* 9,114 of 1,048,575 records displayed(partial results)

Search

Record	ARREST_KEY	ARREST_DATE	PD_DESC	OFNS_DESC	AGE_GROUP	PERP_SEX	PERP_RACE
1	220756993	11/19/2020	RAPE 2	RAPE	18-24	M	BLACK
2	221995093	12/18/2020	SEXUAL ABUSE	SEX CRIMES	45-64	M	WHITE
3	221558560	12/09/2020	RAPE 1	RAPE	<18	M	BLACK
4	217890363	09/15/2020			25-44	M	BLACK
5	219517698	10/21/2020	RAPE 3	RAPE	25-44	M	WHITE HISPANIC
6	218151681	09/21/2020	RAPE 1	RAPE	25-44	M	BLACK
7	174564403	02/08/2018			25-44	M	BLACK
8	215001943	07/07/2020	RAPE 3	RAPE	25-44	M	BLACK
9	215508944	07/20/2020	ARSON 2,3,4	ARSON	<18	M	WHITE
10	217233706	08/31/2020	RAPE 3	RAPE	18-24	M	BLACK
11	217147644	08/28/2020	RAPE 1	RAPE	45-64	M	BLACK
12	215620070	07/22/2020	RAPE 1	RAPE	18-24	M	WHITE
13	212465831	04/28/2020			45-64	M	BLACK HISPANIC
14	216648707	08/16/2020	RAPE 1	RAPE	25-44	M	BLACK

Azure connectivity



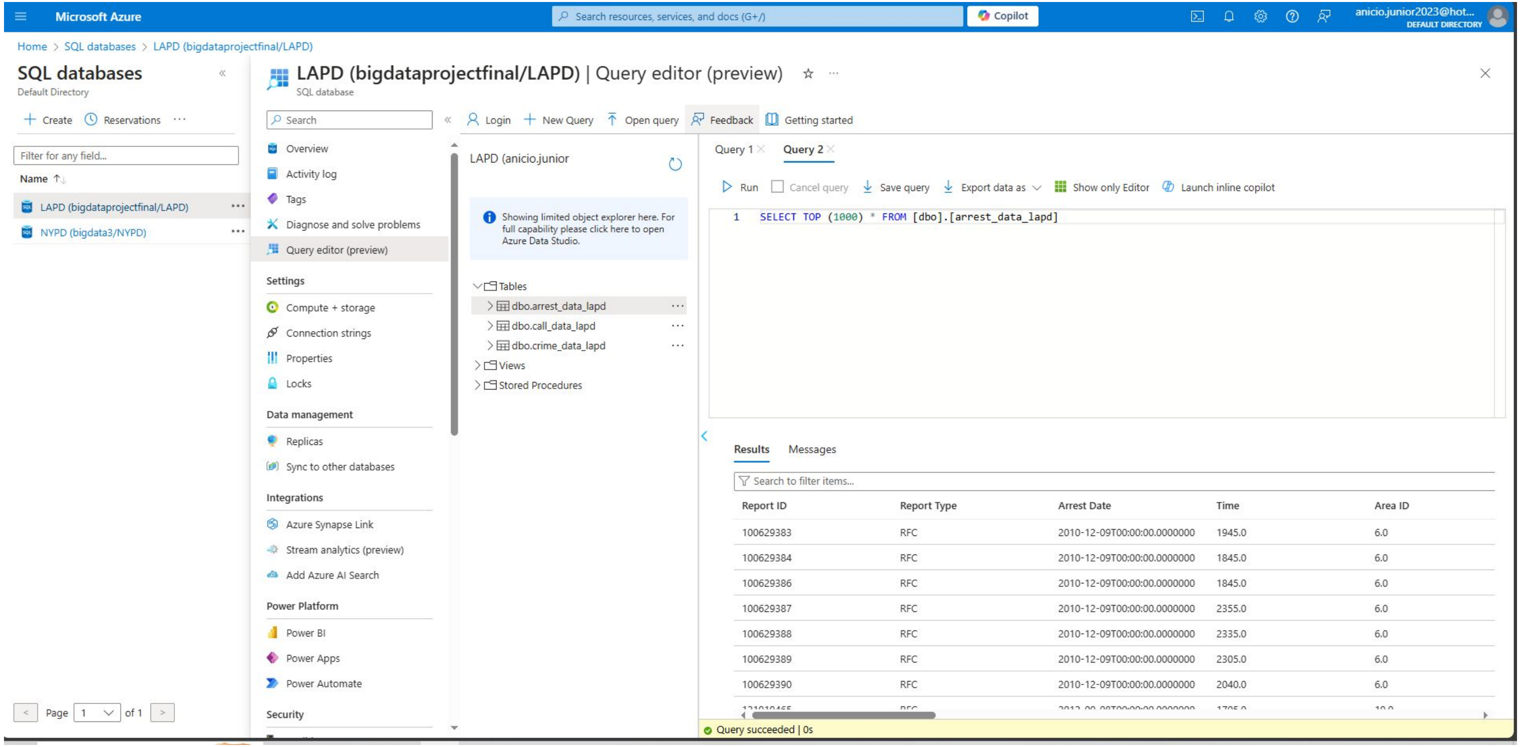
Storage accounts

Default Directory

Showing 1 to 2 of 2 records.

Name	Type	Kind	Resource group	Location	Subscription
lapddata	Storage account	StorageV2	BIGDATA2	Canada Central	Azure subscription 1
nycpddata	Storage account	StorageV2	BIGDATA2	Canada Central	Azure subscription 1

Azure connectivity



Showing limited object explorer here. For full capability please click here to open Azure Data Studio.

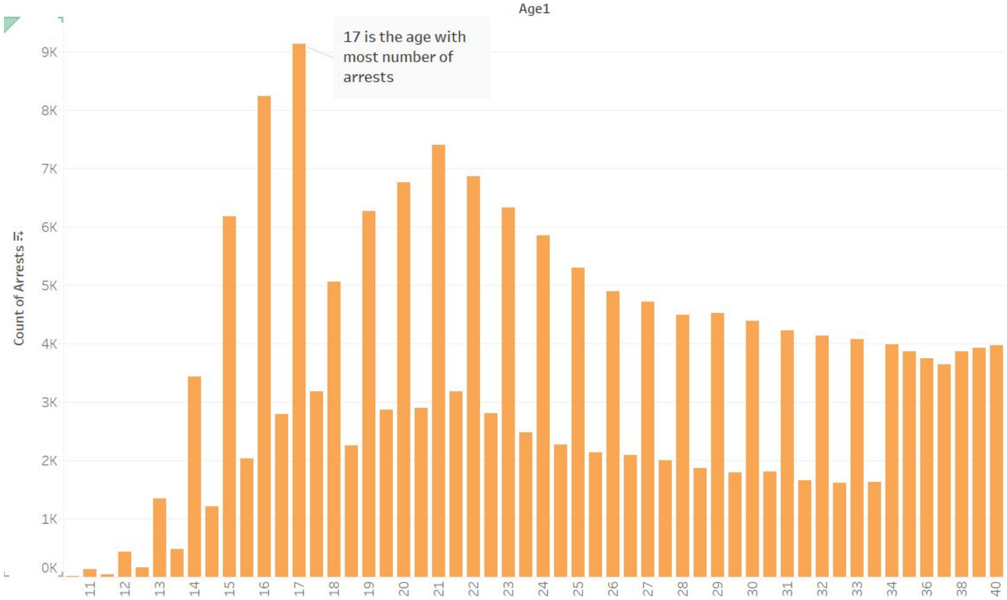
```sql
SELECT TOP (1000) * FROM [dbo].[arrest_data_lapd]
```

Report ID	Report Type	Arrest Date	Time	Area ID
100629383	RFC	2010-12-09T00:00:00.0000000	1945.0	6.0
100629384	RFC	2010-12-09T00:00:00.0000000	1845.0	6.0
100629386	RFC	2010-12-09T00:00:00.0000000	1845.0	6.0
100629387	RFC	2010-12-09T00:00:00.0000000	2355.0	6.0
100629388	RFC	2010-12-09T00:00:00.0000000	2335.0	6.0
100629389	RFC	2010-12-09T00:00:00.0000000	2305.0	6.0
100629390	RFC	2010-12-09T00:00:00.0000000	2040.0	6.0

Query succeeded | 0s

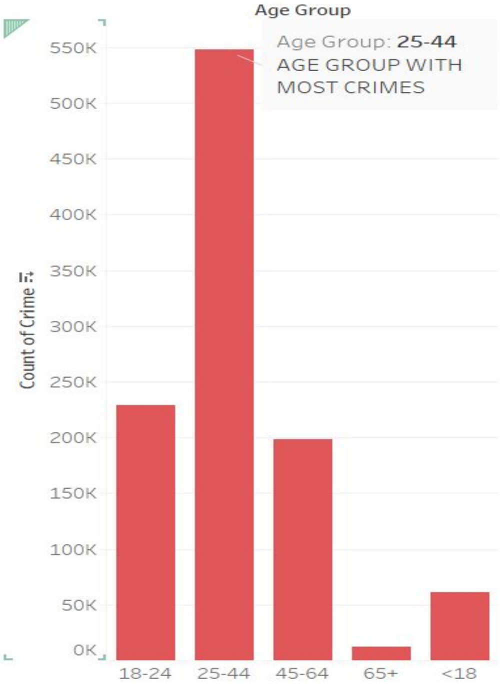
Results and graphs

CRIME VS AGE (LAPD)



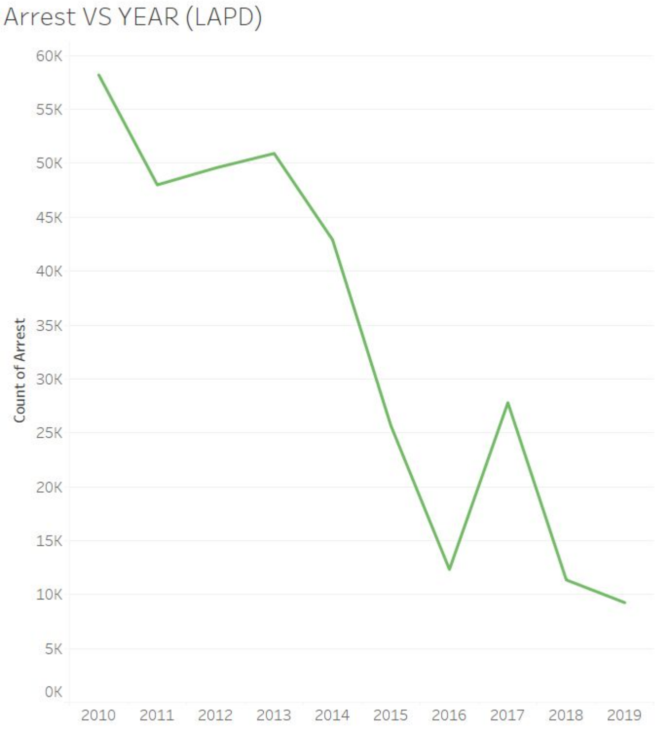
17 is the age with most number of arrests

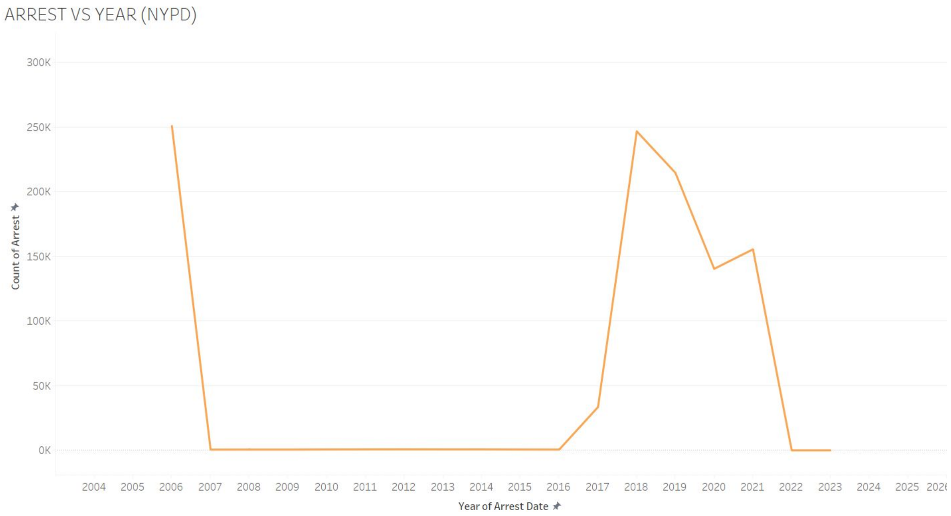
Crime vs AGE NYPD



Age Group: **25-44**
AGE GROUP WITH MOST CRIMES

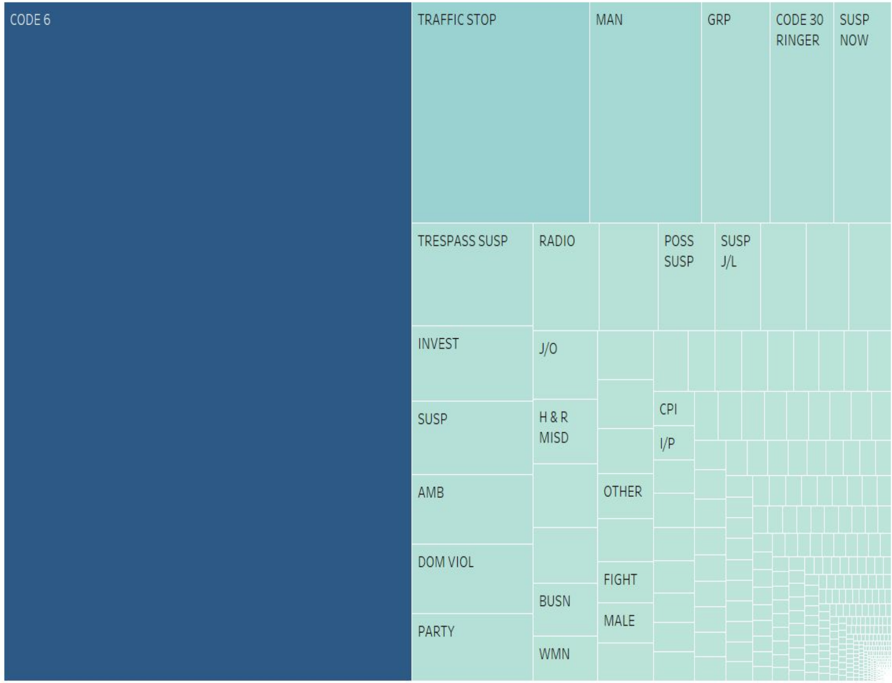
Results and graphs



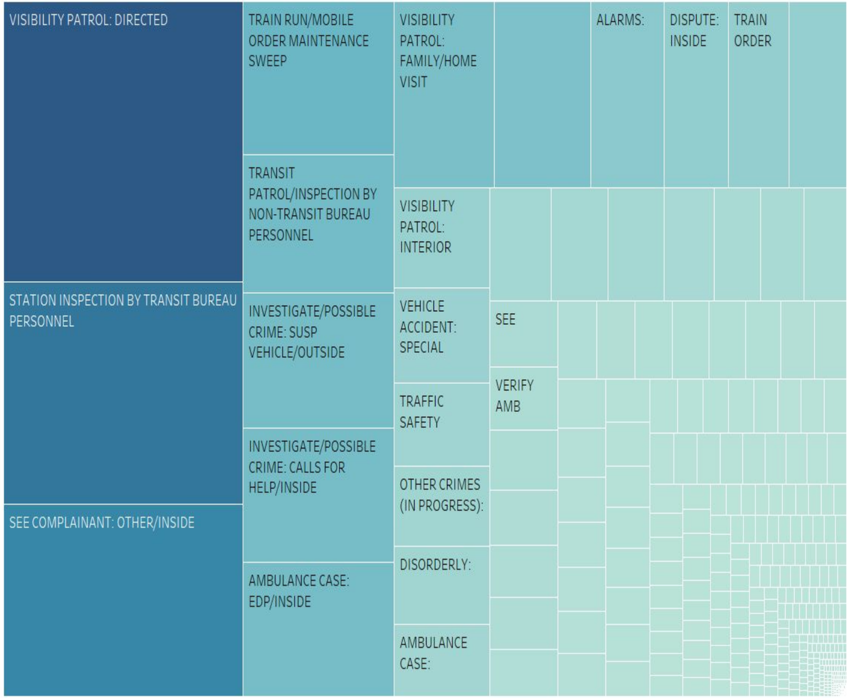


Results and graphs

Most Calls For Service (LAPD)

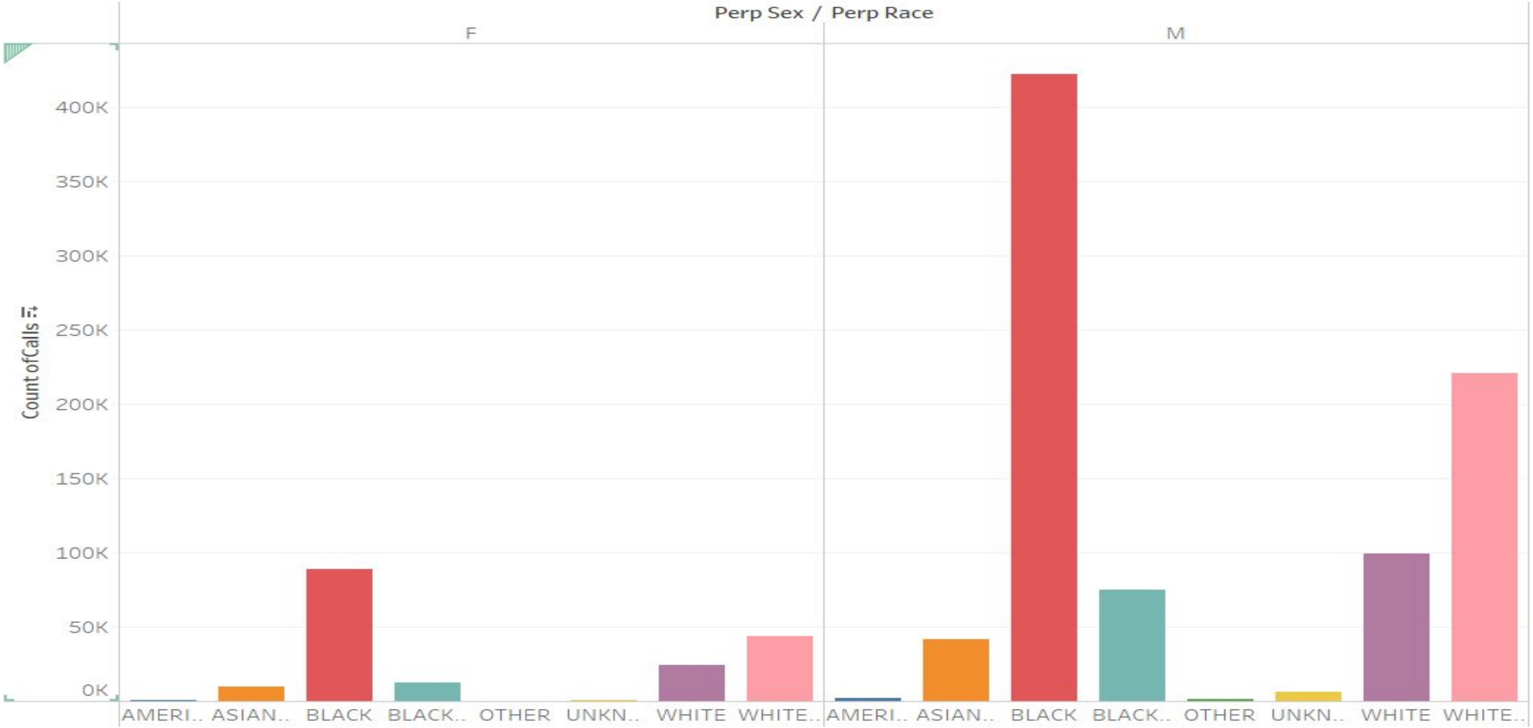


Most CALLS for Service(NYPD)



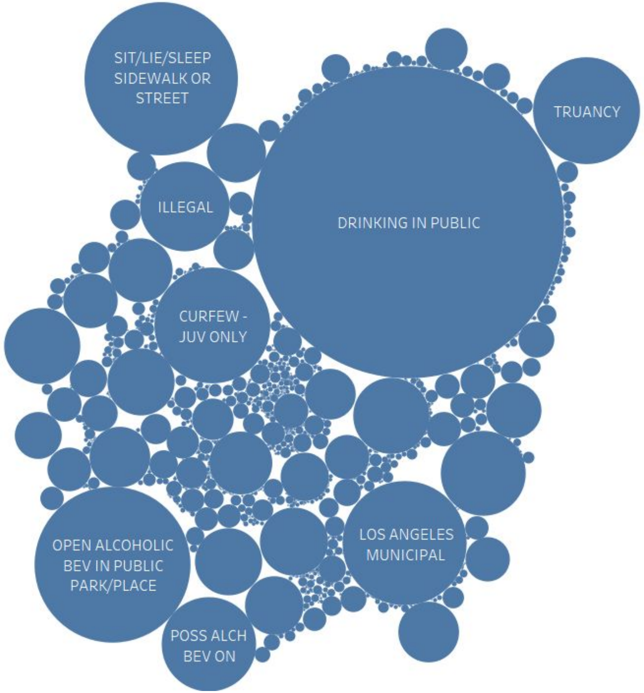
Results and graphs

NUMBER OF CALLS FOR CRIME VS PERSON (SEX/ RACE)

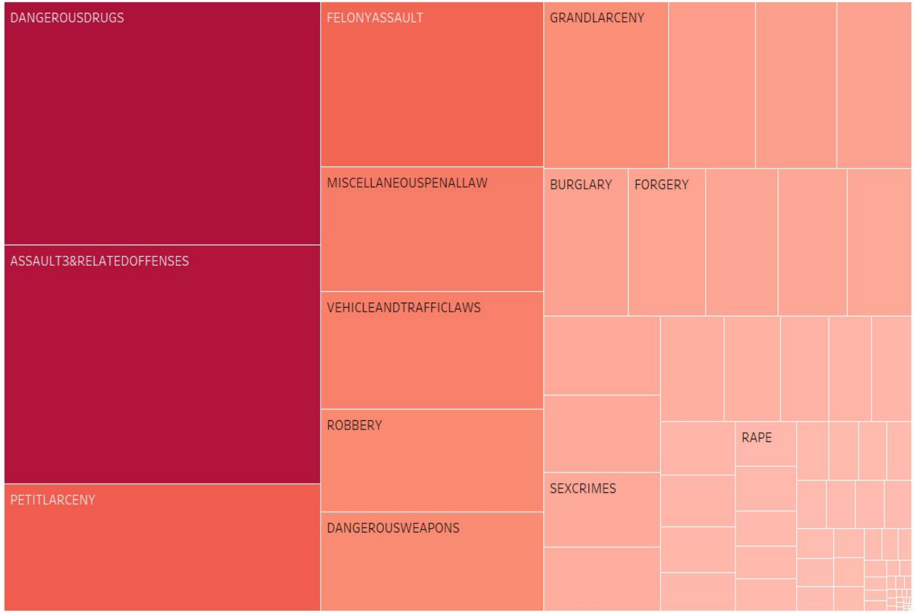


Results and graphs

MOST CRIMES (LAPD)



MOST CRIMES(NYPD)



Conclusion

Key Findings:

- **Crime vs. Age:**
  - **LAPD:** Peak arrests at age 17, smaller peak around ages 23-24.
  - **NYPD:** Highest crime count in the 25-44 age group.
- **Arrests vs. Year:**
  - **LAPD:** Overall declining trend from 2010 to 2019.
  - **NYPD:** Significant drop around 2012, followed by fluctuations.
- **Most Calls for Service:**
  - **LAPD:** 'Code 6' is the most frequent call.
  - **NYPD:** 'Visibility Patrol Directed' is the most common call.

Implications:

- Differences in age-related crime trends suggest targeted interventions.
- Yearly arrest trends indicate varying impacts of policies and socio-economic factors.
- Most common service calls reflect different policing priorities and strategies.

Conclusion:

- Understanding these patterns can help optimize law enforcement strategies and improve crime prevention efforts.
- Insights are valuable for developing targeted interventions in both cities.