Name:Amruta Naik

NUID:002891105

Individual Project Part 1

ER STUDIO:

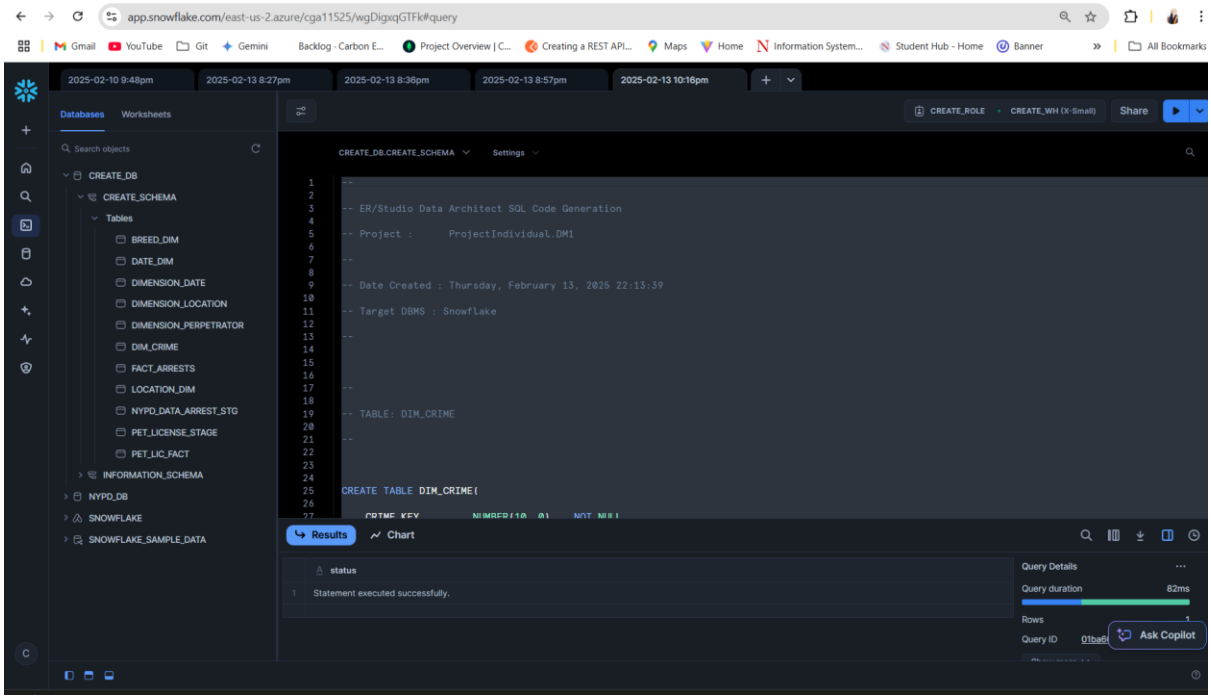
Filename:ER-STUDIO: ProjectIndividualAmruta

ProjectIndividualAmruta.DM1

SQLscripts ran to create tables in snowflake

DimensionalModellingfile.sql

Screen of tables Created on Snowflake:



Alteryx Profiling

Alteryx is used for profiling and also cleaning the data

finalprojectalteryx.yxmd

Problems identified:

1. Used basic profiling and field summary to identify the issues in the NYPD data arrest file.
2. Identified the Issues and then cleaned the data using Alteryx.

List of issues identified while profiling

1.Used select in Alteryx to Change all the data types as per the meta data for all 22 columns.

2.Changed the ARREST_DATE to appropriate snowflake acceptable format YYYY-MM-DD saved this with DateTimeNow

3.Identified all the null values and replaced all the Numeric values with 0 and string valued are replaced by known Bu using functions in ALTERYX

Identified nulls in

- PD_CD numeric
- PD_DESC string
- KY_CD numeric
- OFNS_DESC string
- LAW_CODE string
- LAW_CAT_CD had 9,I and null values which was handled

4.For latitude longitude and X-coord and Y -coord

Where ever there is 0 or null values. I calculated the mean and then replaced 0 or null values with the Mean for all of the 4 columns.

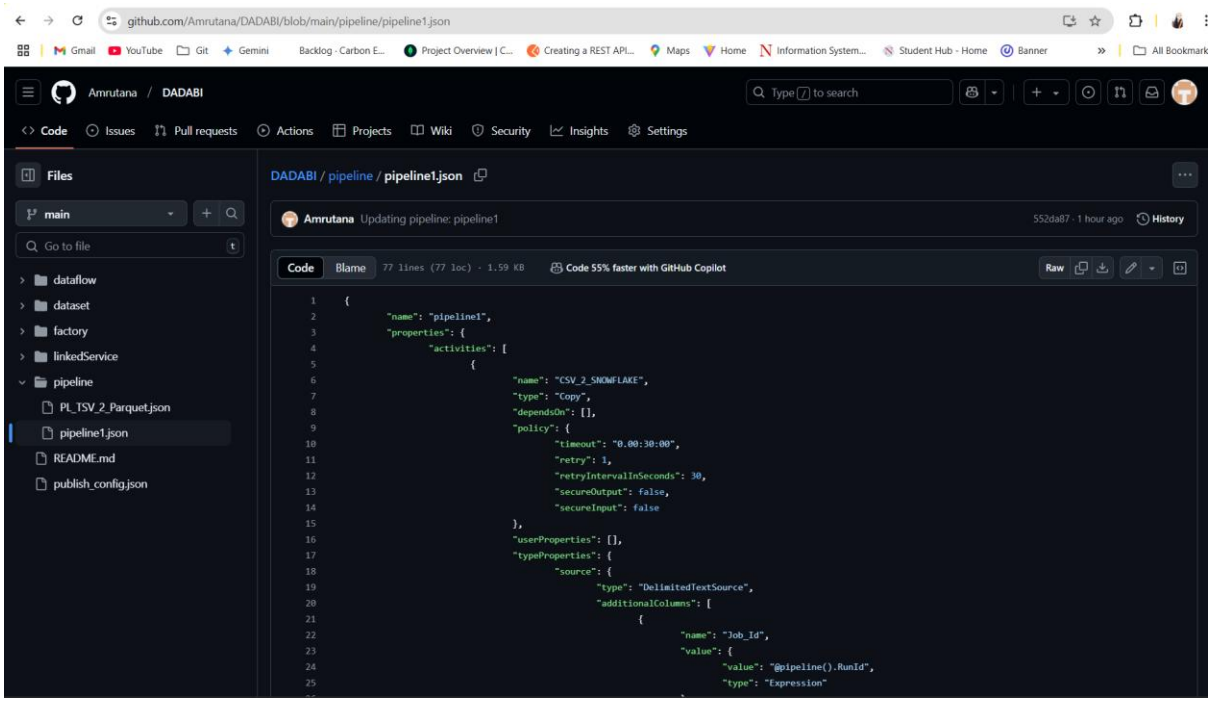
Once profiling is done I stored my data locally and then used CSV to snowkflake pipelining and did a stage table in snowflake

Created a stage table in snowflake with

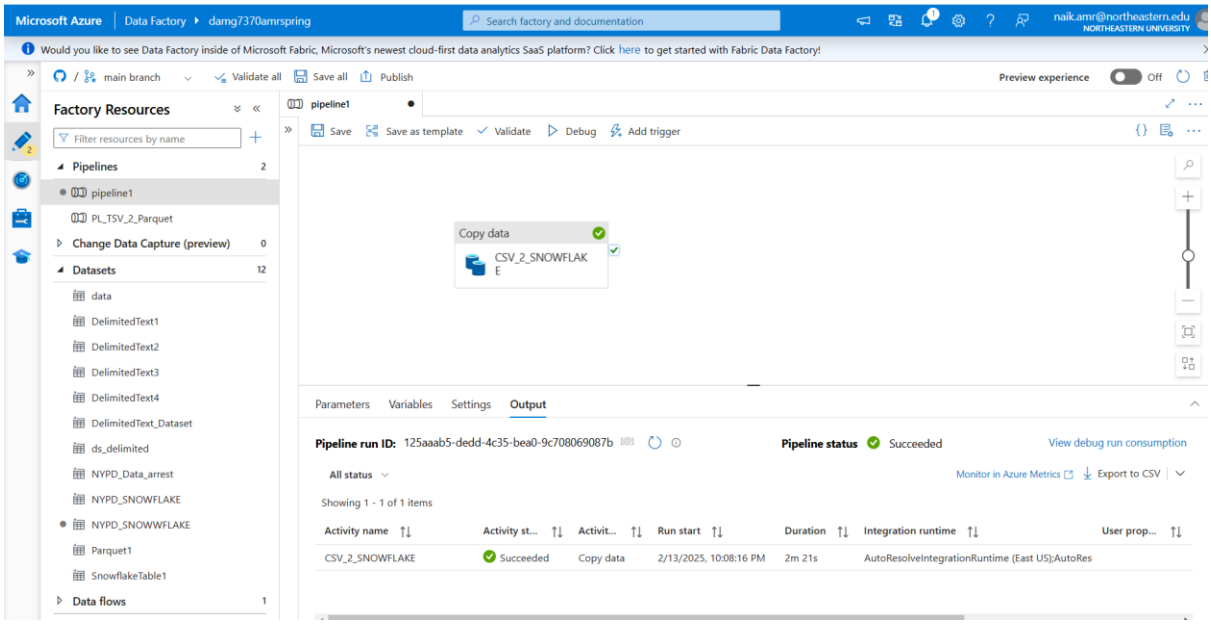
```
CREATE OR REPLACE TABLE NYPD_ARREST_SCHEMA.arrest_stg (
    ARREST_KEY VARCHAR(16384),
    ARREST_DATE VARCHAR(16384),
    PD_CD INT,
    PD_DESC VARCHAR(16384),
    KY_CD INT,
    OFNS_DESC VARCHAR(16384),
    LAW_CODE VARCHAR(16384),
    LAW_CAT_CD VARCHAR(16384),
    ARREST_BORO VARCHAR(16384),
    ARREST_PRECINCT INT,
    JURISDICTION_CODE BIGINT,
    AGE_GROUP VARCHAR(16384),
    PERP_SEX VARCHAR(16384),
    PERP_RACE VARCHAR(16384),
    X_COORD_CD DOUBLE,
    Y_COORD_CD DOUBLE,
    Latitude DOUBLE,
    Longitude DOUBLE,
    NEW_GEOREFERENCED VARCHAR(16384),
    DateTime_Out DATE,
    LOAD_DATE DATE,
    JOB_ID VARCHAR(50)
);
```

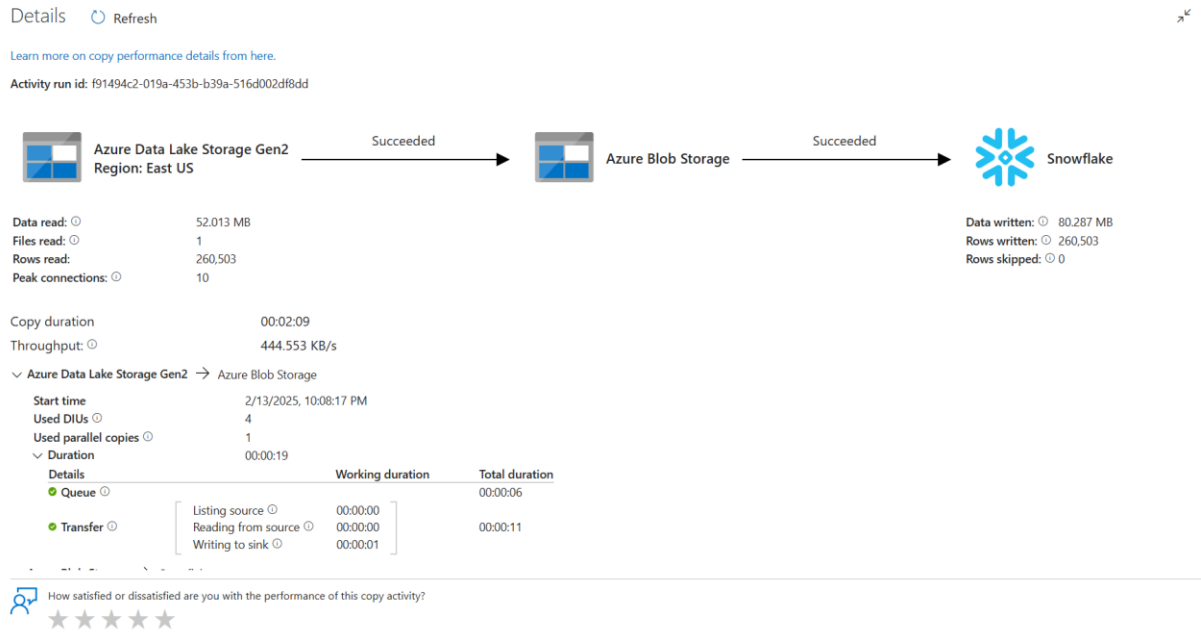
GIT repository Link

https://github.com/Amrutana/DADABI/tree/main/pipeline

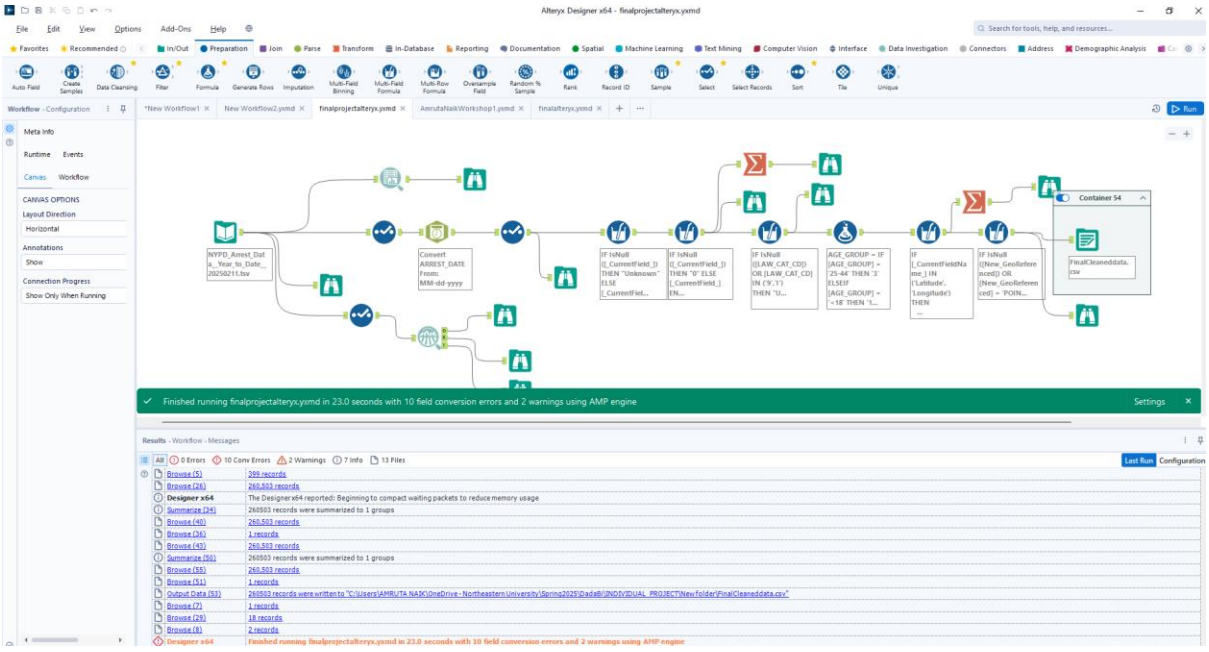


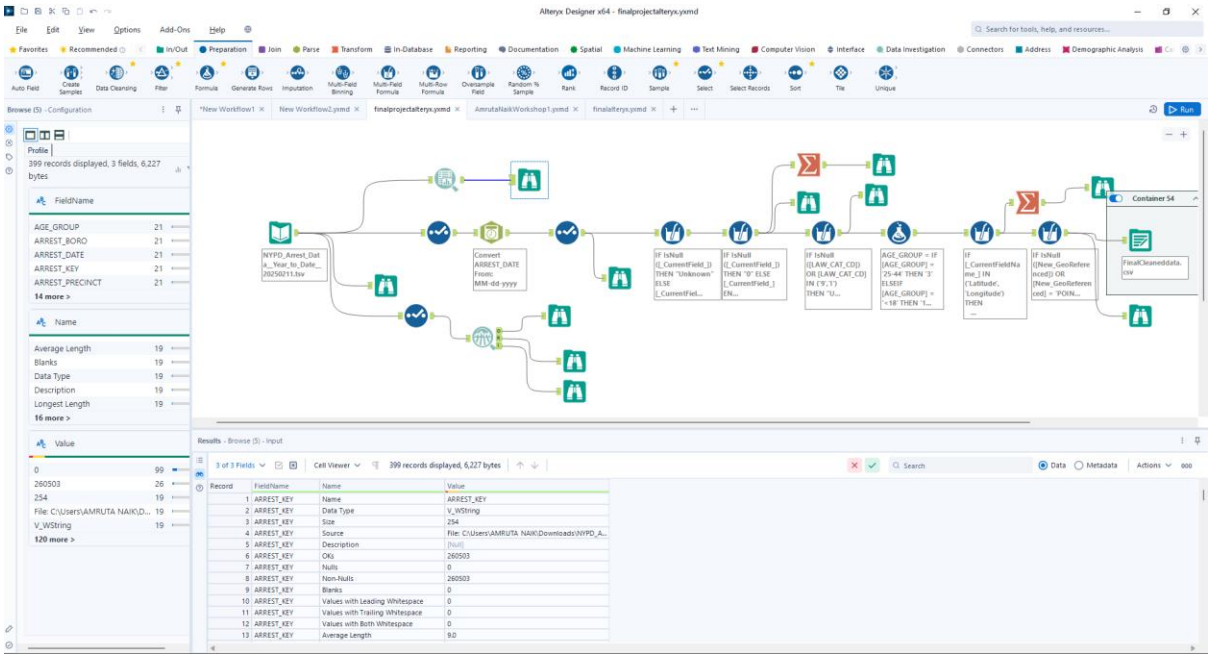
ADF pipeline Screenshot:

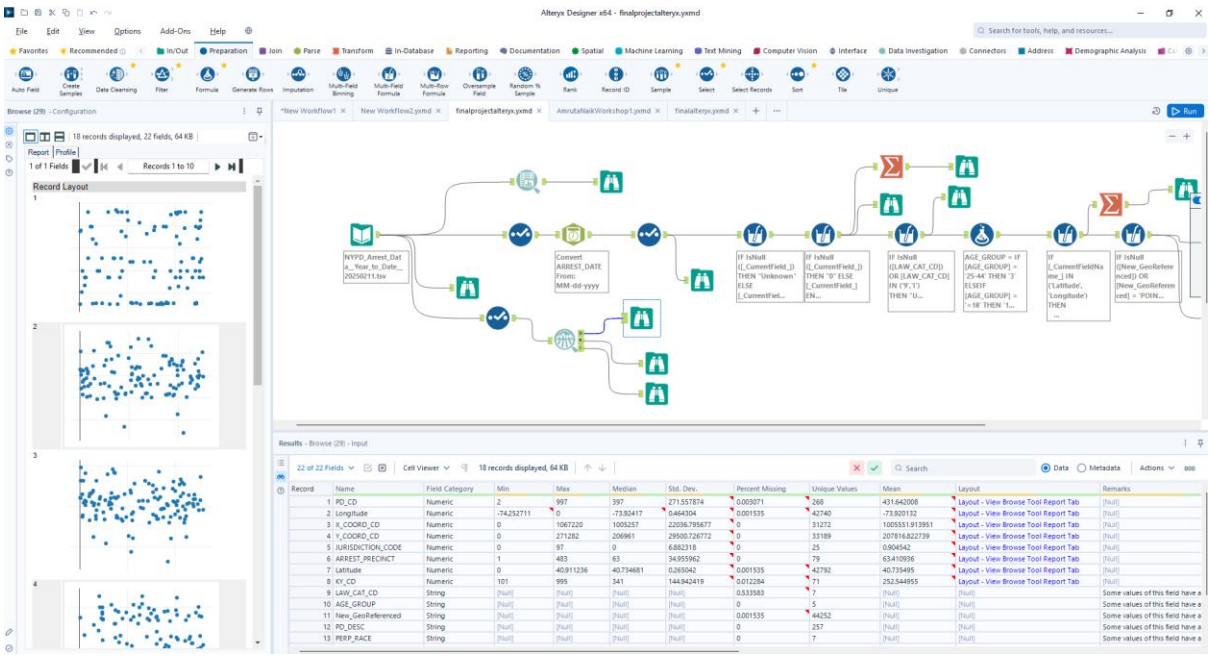


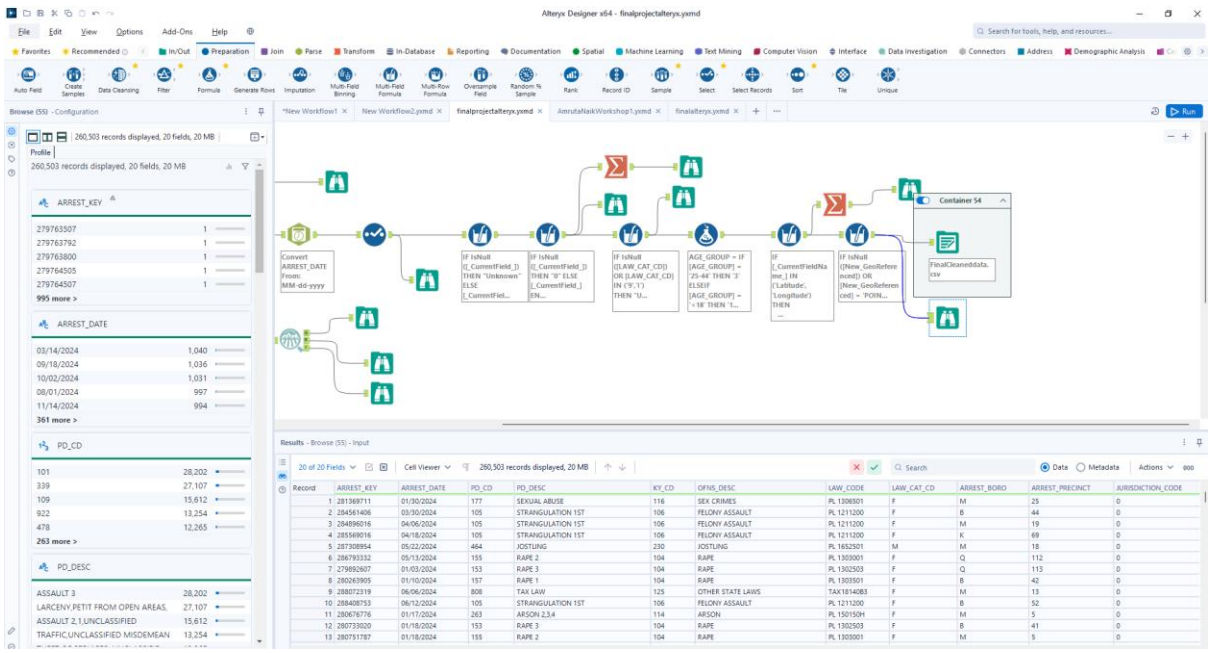


Alteryx workflow run




Alteryx Designer x64 - finalprojectalteryx.yxmd
File Edit View Options Add-Ons Help
Search for tools, help, and resources...
Favorites Recommended In/Out Preparation Join Parse Transform In-Database Reporting Documentation Spatial Machine Learning Text Mining Computer Vision Interface Data Investigation Connectors Address Demographic Analysis
Auto Field Create Samples Data Cleansing Filter Formula Generate Rows Imputation Multi-Field Binning Multi-Field Formula Multi-Row Formula Oversample Field Random % Sample Rank Record ID Sample Select Select Records Sort Tile Unique
Browse (5) - Configuration
*New Workflow1 New Workflow2.yxmd finalprojectalteryx.yxmd AmrutaNaikWorkshop1.yxmd finalalteryx.yxmd
Run
Profile
399 records displayed, 3 fields, 6,227 bytes
FieldName
AGE_GROUP 21
ARREST_BORO 21
ARREST_DATE 21
ARREST_KEY 21
ARREST_PRECINCT 21
14 more >
Name
Average Length 19
Blanks 19
Data Type 19
Description 19
Longest Length 19
16 more >
Value
0 99
260503 26
254 19
File: C:\Users\AMRUTA NAIK\D... 19
V_WString 19
120 more >
NYPD_Arrest_Data_Year_to_Date_20250211.tsv
Convert ARREST_DATE From: MM-dd-yyyy
IF IsNull([_CurrentField_]) THEN 'Unknown' ELSE [_CurrentFiel...
IF IsNull([_CurrentField_]) THEN '0' ELSE [_CurrentField_] EN...
IF IsNull([LAW_CAT_CD]) OR [LAW_CAT_CD] IN ('F','I') THEN 'U...
AGE_GROUP = IF [AGE_GROUP] = '25-44' THEN '3' ELSEIF [AGE_GROUP] = '<18' THEN '1'...
IF [_CurrentFieldName_] IN ('Latitude', 'Longitude') THEN ...
IF IsNull([New_GeoReferenced]) OR [New_GeoReferenced] = 'POIN...
Container 54
FinalCleaneddata.csv
Results - Browse (5) - Input
3 of 3 Fields Cell Viewer 399 records displayed, 6,227 bytes
Search Data Metadata Actions
Record FieldName Name Value
1 ARREST_KEY Name ARREST_KEY
2 ARREST_KEY Data Type V_WString
3 ARREST_KEY Size 254
4 ARREST_KEY Source File: C:\Users\AMRUTA NAIK\Downloads\NYPD_A...
5 ARREST_KEY Description [Null]
6 ARREST_KEY OKs 260503
7 ARREST_KEY Nulls 0
8 ARREST_KEY Non-Nulls 260503
9 ARREST_KEY Blanks 0
10 ARREST_KEY Values with Leading Whitespace 0
11 ARREST_KEY Values with Trailing Whitespace 0
12 ARREST_KEY Values with Both Whitespace 0
13 ARREST_KEY Average Length 9.0

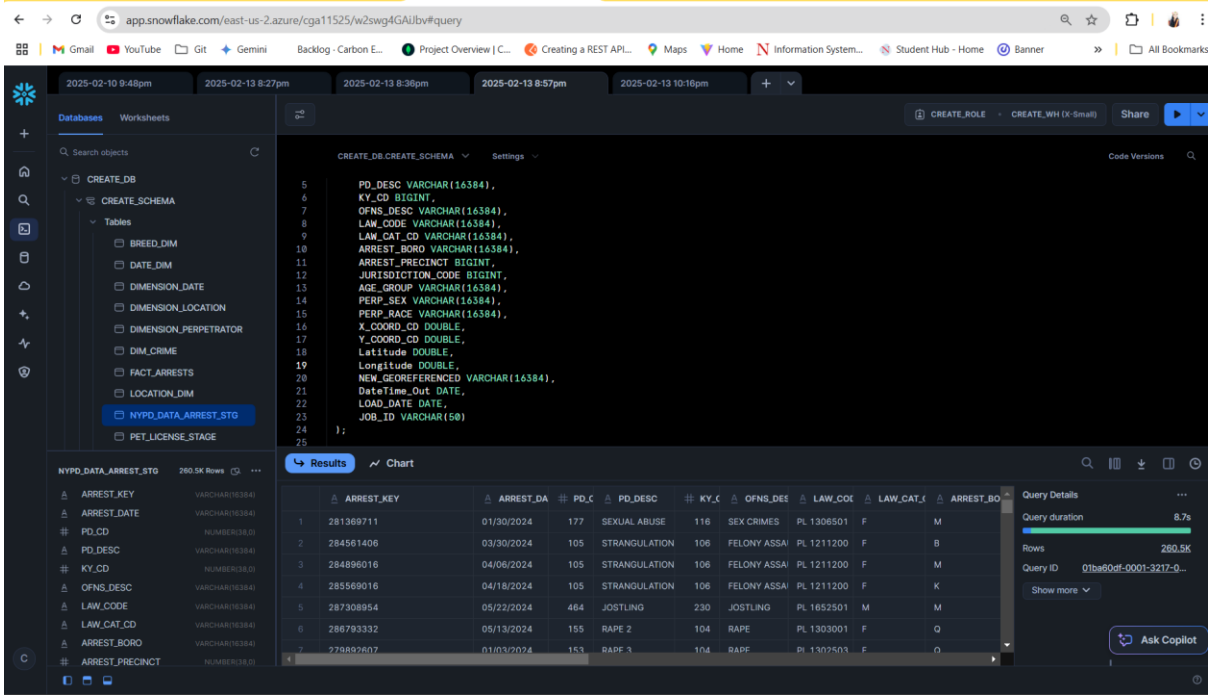

Alteryx Designer x64 - finalprojectalteryx.yxmd
File Edit View Options Add-Ons Help
Search for tools, help, and resources...
Favorites Recommended In/Out Preparation Join Parse Transform In-Database Reporting Documentation Spatial Machine Learning Text Mining Computer Vision Interface Data Investigation Connectors Address Demographic Analysis
Auto Field Create Samples Data Cleansing Filter Formula Generate Rows Imputation Multi-Field Binning Multi-Field Formula Multi-Row Formula Oversample Field Random % Sample Rank Record ID Sample Select Select Records Sort Tile Unique
Browse (29) - Configuration
*New Workflow1 New Workflow2.yxmd finalprojectalteryx.yxmd AmrutaNaikWorkshop1.yxmd finalalteryx.yxmd
Run
18 records displayed, 22 fields, 64 KB
Report Profile
1 of 1 Fields Records 1 to 10
Record Layout
NYPD_Arrest_Data_Year_to_Date_20250211.tsv
Convert ARREST_DATE From: MM-dd-yyyy
IF IsNull([_CurrentField_]) THEN 'Unknown' ELSE [_CurrentFiel...
IF IsNull([_CurrentField_]) THEN '0' ELSE [_CurrentField_] EN...
IF IsNull([LAW_CAT_CD]) OR [LAW_CAT_CD] IN ('F','I') THEN 'U...
AGE_GROUP = IF [AGE_GROUP] = '25-44' THEN '3' ELSEIF [AGE_GROUP] = '<18' THEN '1'...
IF [_CurrentFieldName_] IN ('Latitude', 'Longitude') THEN ...
IF IsNull([New_GeoReferenced]) OR [New_GeoReferenced] = 'POIN...
Results - Browse (29) - Input
22 of 22 Fields Cell Viewer 18 records displayed, 64 KB
Search Data Metadata Actions
Record Name Field Category Min Max Median Std. Dev. Percent Missing Unique Values Mean Layout Remarks
1 PD_CD Numeric 2 997 397 271.557874 0.003071 268 431.642008 Layout - View Browse Tool Report Tab [Null]
2 Longitude Numeric -74.252711 0 -73.92417 0.464304 0.001535 42740 -73.920132 Layout - View Browse Tool Report Tab [Null]
3 X_COORD_CD Numeric 0 1067220 1005257 22036.795677 0 31272 1005531.913951 Layout - View Browse Tool Report Tab [Null]
4 Y_COORD_CD Numeric 0 271282 206961 29500.726772 0 33189 207816.822739 Layout - View Browse Tool Report Tab [Null]
5 JURISDICTION_CODE Numeric 0 97 0 6.882318 0 25 0.904542 Layout - View Browse Tool Report Tab [Null]
6 ARREST_PRECINCT Numeric 1 483 63 34.955962 0 79 63.410936 Layout - View Browse Tool Report Tab [Null]
7 Latitude Numeric 0 40.911236 40.734681 0.265042 0.001535 42792 40.739495 Layout - View Browse Tool Report Tab [Null]
8 KY_CD Numeric 101 995 341 144.942419 0.012284 71 252.544955 Layout - View Browse Tool Report Tab [Null]
9 LAW_CAT_CD String [Null] [Null] [Null] [Null] 0.533583 7 [Null] [Null] Some values of this field have a
10 AGE_GROUP String [Null] [Null] [Null] [Null] 0 5 [Null] [Null] Some values of this field have a
11 New_GeoReferenced String [Null] [Null] [Null] [Null] 0.001535 44252 [Null] [Null] Some values of this field have a
12 PD_DESC String [Null] [Null] [Null] [Null] 0 257 [Null] [Null] Some values of this field have a
13 PERP_RACE String [Null] [Null] [Null] [Null] 0 7 [Null] [Null] Some values of this field have a



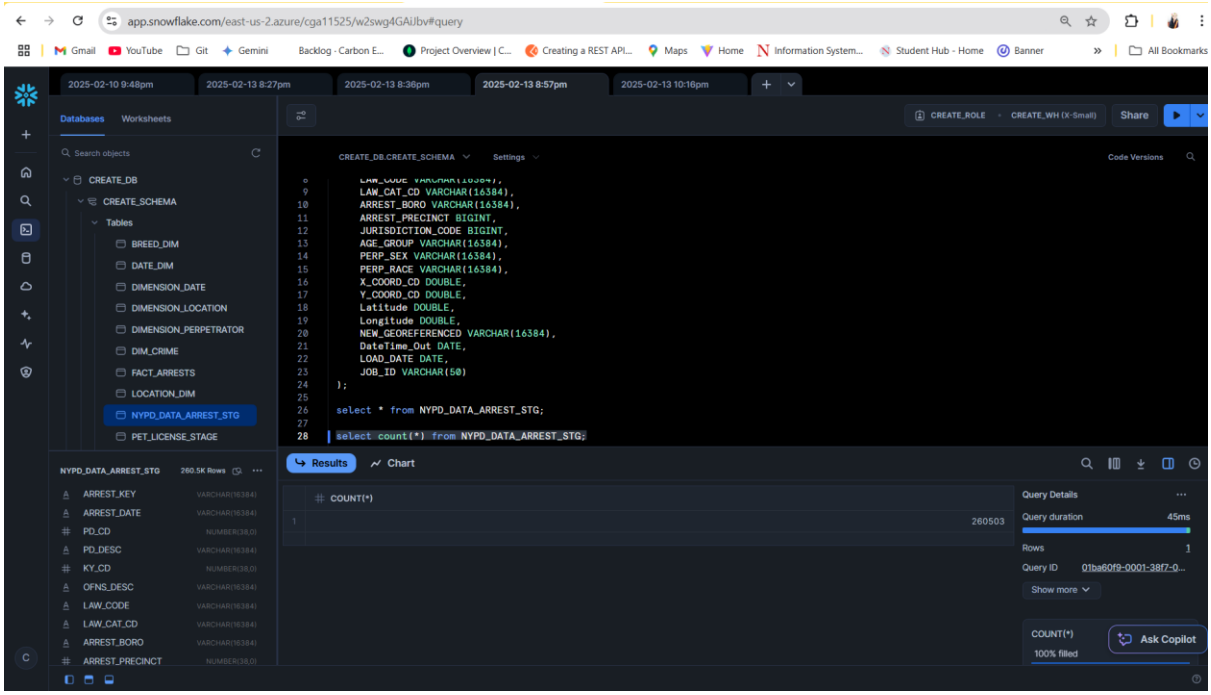
Output csv file of cleaned data

FinalCleaneddata.csv

Data Populated in Snowflake after running the pipeline



- Row counts using SQL query



Running SQL scripts as per the business requirements

1. How many arrests occurred on any specific day, week, month, quarter, or year?

-- Daily arrests

SELECT

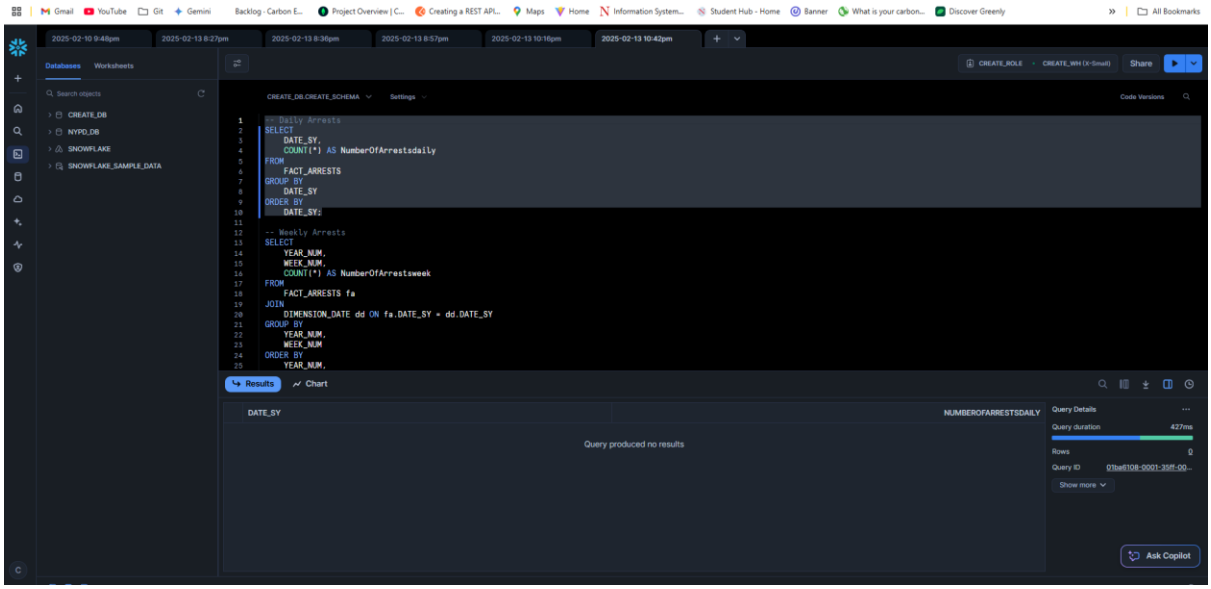
DATE_SY,

COUNT(*) **AS** ArrestCount

FROM FACT_ARRESTS

GROUP BY DATE_SY

ORDER BY DATE_SY;



-- Weekly arrests

SELECT

YEAR(DATE_SY) **AS** ArrestYear,

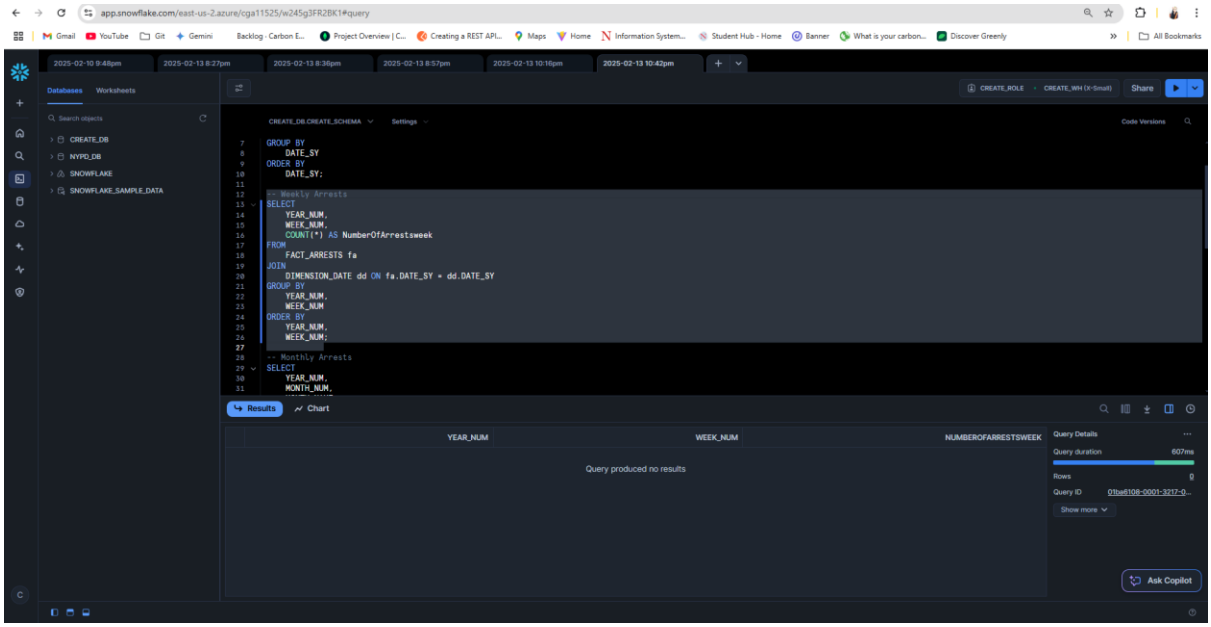
DATEPART(wk, DATE_SY) **AS** ArrestWeek,

COUNT(*) **AS** ArrestCount

FROM FACT_ARRESTS

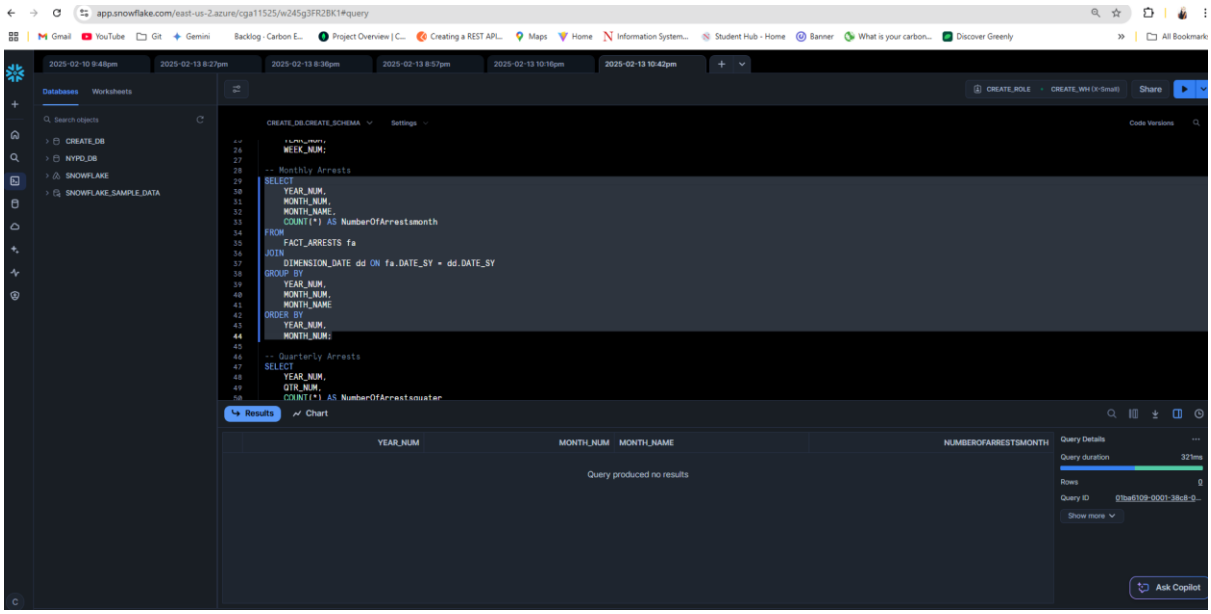
GROUP BY YEAR(DATE_SY), DATEPART(wk, DATE_SY)

ORDER BY YEAR(DATE_SY), DATEPART(wk, DATE_SY);



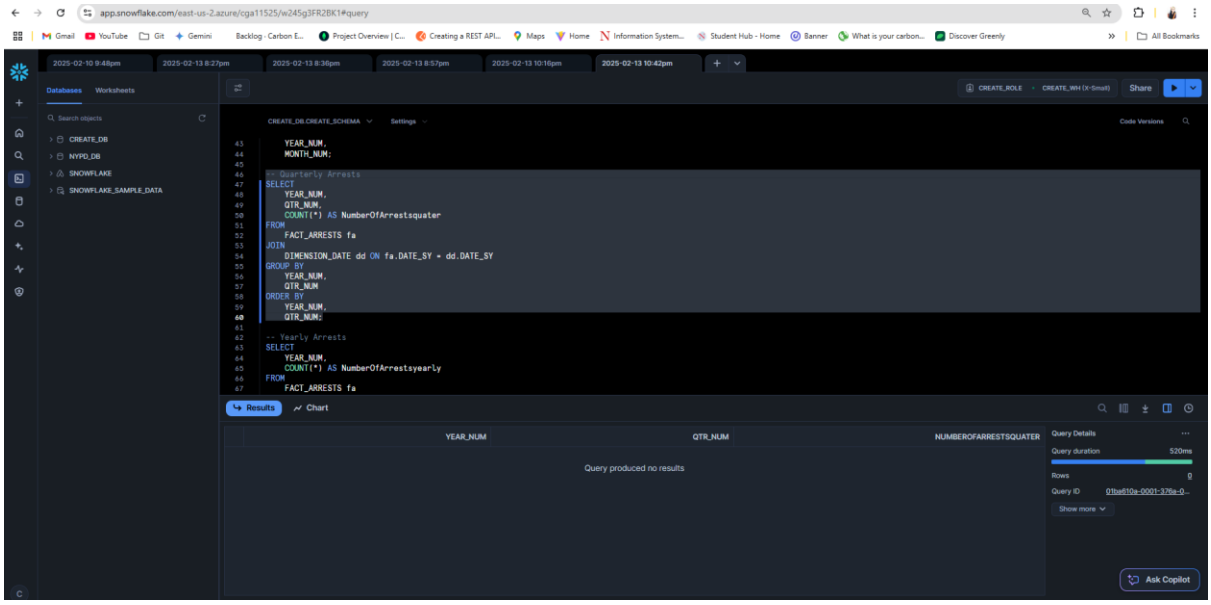
-- Monthly arrests

```
SELECT
    YEAR(DATE_SY) AS ArrestYear,
    MONTH(DATE_SY) AS ArrestMonth,
    COUNT(*) AS ArrestCount
FROM FACT_ARRESTS
GROUP BY YEAR(DATE_SY), MONTH(DATE_SY)
ORDER BY YEAR(DATE_SY), MONTH(DATE_SY);
```



-- Quarterly arrests

```
SELECT
    YEAR(DATE_SY) AS ArrestYear,
    DATEPART(qq, DATE_SY) AS ArrestQuarter,
    COUNT(*) AS ArrestCount
FROM FACT_ARRESTS
GROUP BY YEAR(DATE_SY), DATEPART(qq, DATE_SY)
ORDER BY YEAR(DATE_SY), DATEPART(qq, DATE_SY);
```



-- Yearly arrests

SELECT

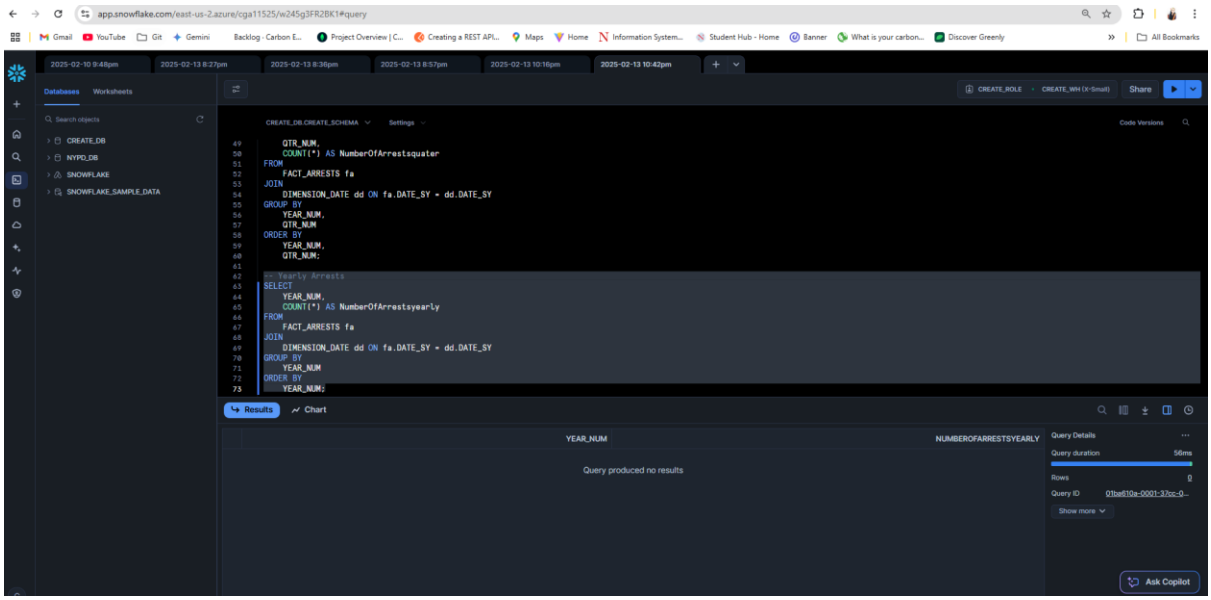
YEAR(DATE_SY) **AS** ArrestYear,

COUNT(*) **AS** ArrestCount

FROM FACT_ARRESTS

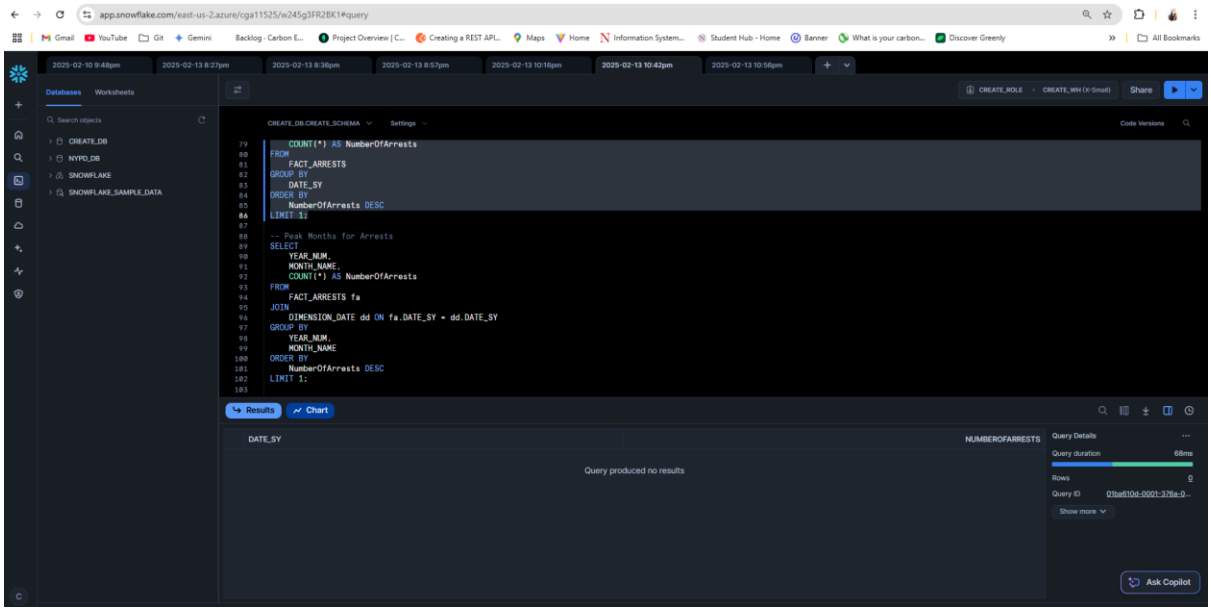
GROUP BY YEAR(DATE_SY)

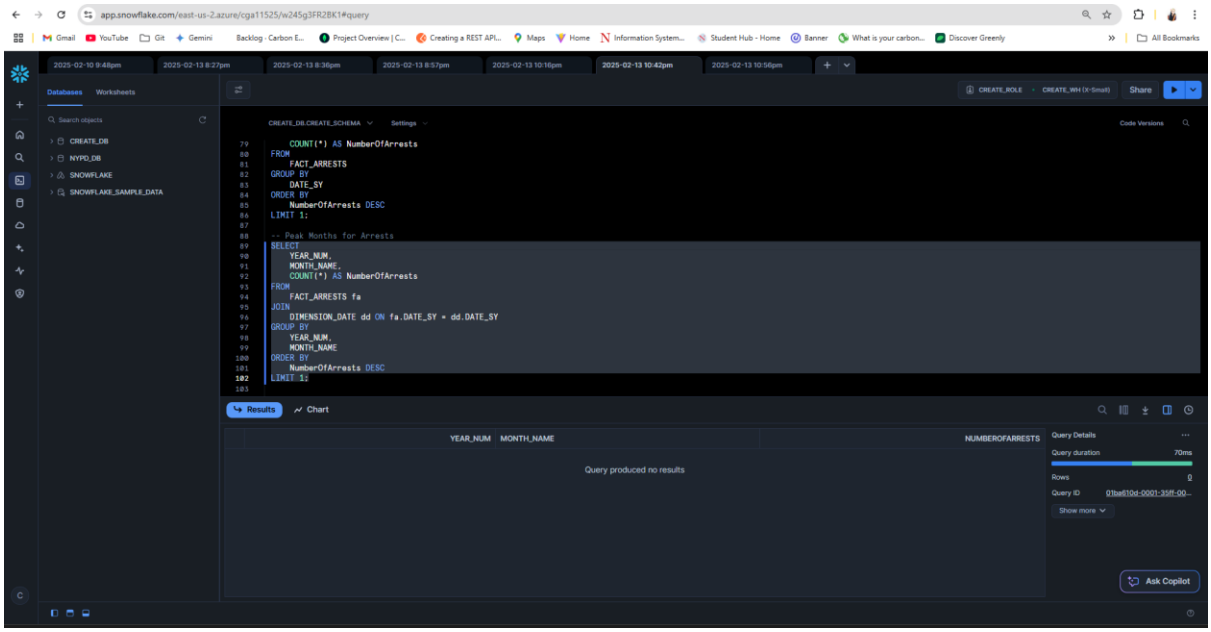
ORDER BY YEAR(DATE_SY);



2. What are the peak days and months for arrests?

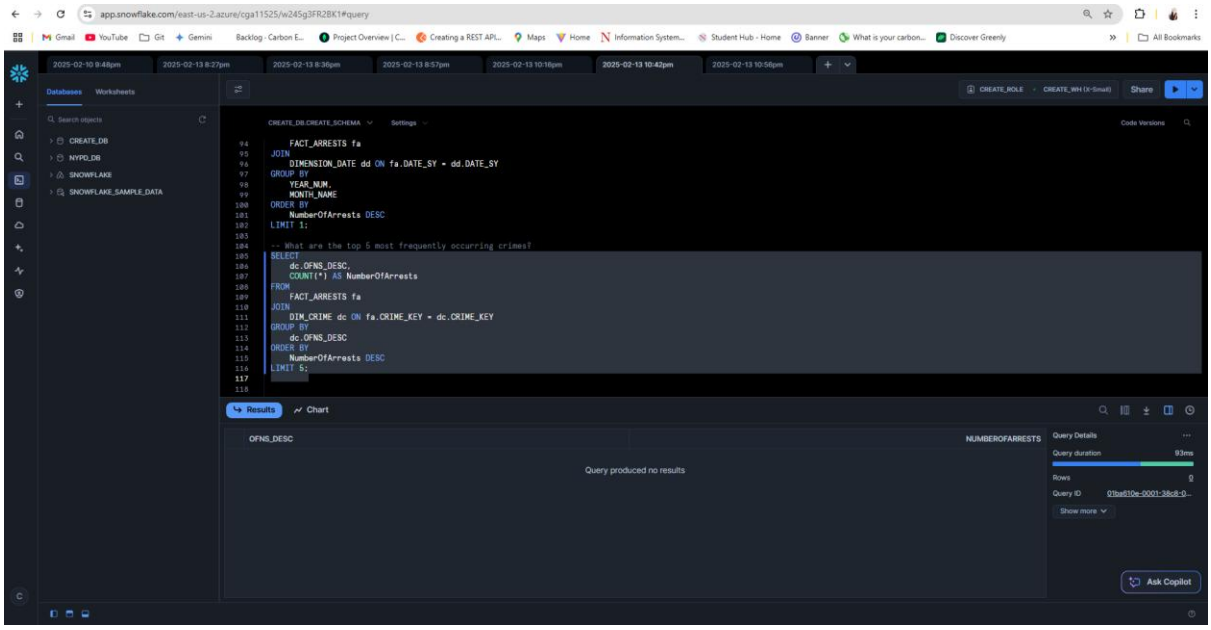
```
-- Peak Days for Arrests
SELECT
  DATE_SY,
  COUNT(*) AS NumberOfArrests
FROM
  FACT_ARRESTS
GROUP BY
  DATE_SY
ORDER BY
  NumberOfArrests DESC
LIMIT 1;
-- Peak Months for Arrests
SELECT
  YEAR_NUM,
  MONTH_NAME,
  COUNT(*) AS NumberOfArrests
FROM
  FACT_ARRESTS fa
JOIN
  DIMENSION_DATE dd ON fa.DATE_SY = dd.DATE_SY
GROUP BY
  YEAR_NUM,
  MONTH_NAME
ORDER BY
  NumberOfArrests DESC
LIMIT 1;
```





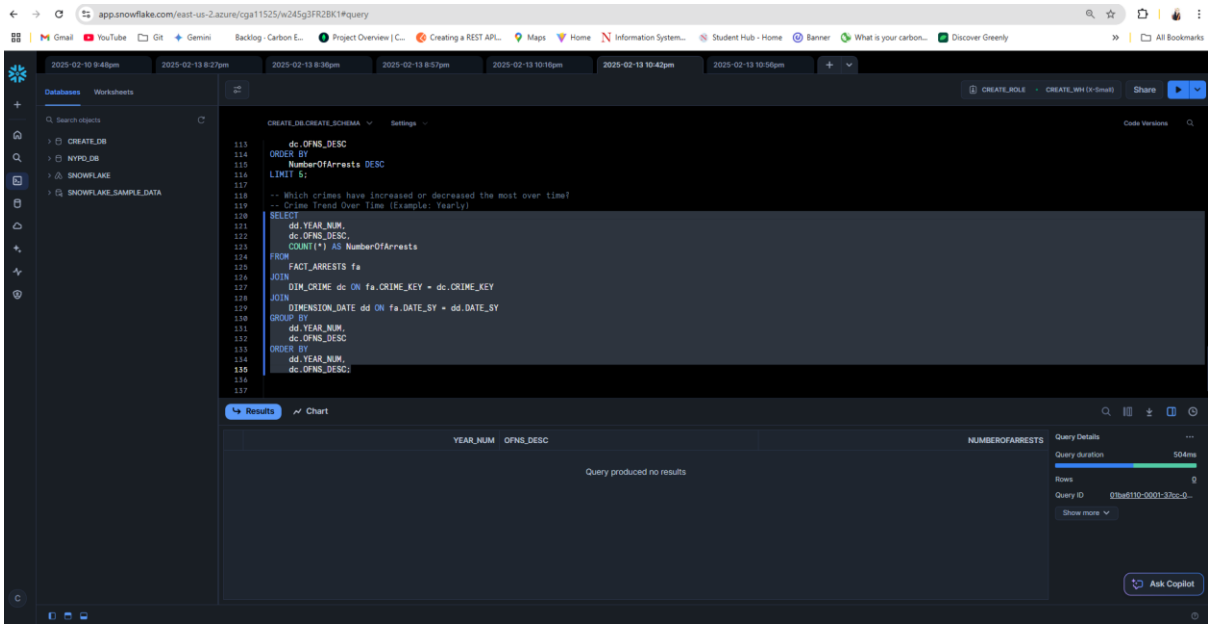
3. What are the top 5 most frequently occurring crimes?

```
SELECT
   dc.OFNS_DESC,
   COUNT(*) AS NumberOfArrests
FROM
   FACT_ARRESTS fa
JOIN
   DIM_CRIME dc ON fa.CRIME_KEY = dc.CRIME_KEY
GROUP BY
   dc.OFNS_DESC
ORDER BY
   NumberOfArrests DESC
LIMIT 5;
```



4. Which crimes have increased or decreased the most over time?

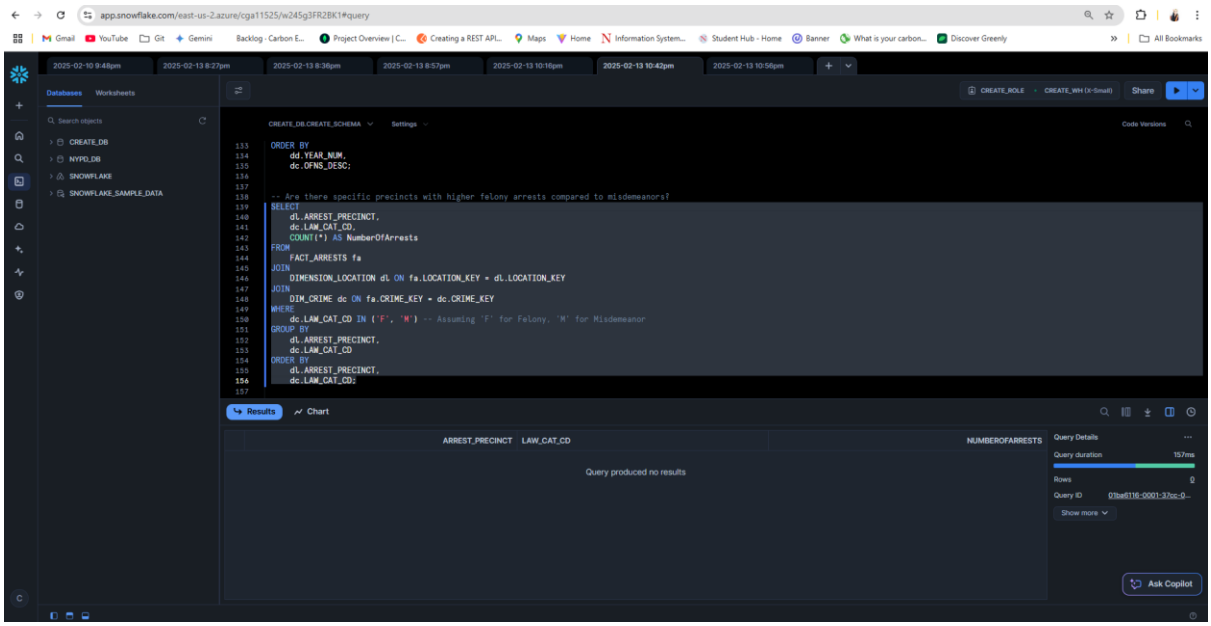
```
SELECT
   dd.YEAR_NUM,
   dc.OFNS_DESC,
   COUNT(*) AS NumberOfArrests
FROM
   FACT_ARRESTS fa
JOIN
   DIM_CRIME dc ON fa.CRIME_KEY = dc.CRIME_KEY
JOIN
   DIMENSION_DATE dd ON fa.DATE_SY = dd.DATE_SY
GROUP BY
   dd.YEAR_NUM,
   dc.OFNS_DESC
ORDER BY
   dd.YEAR_NUM,
   dc.OFNS_DESC;
```



5. *Are there specific precincts with higher felony arrests compared to misdemeanors?*

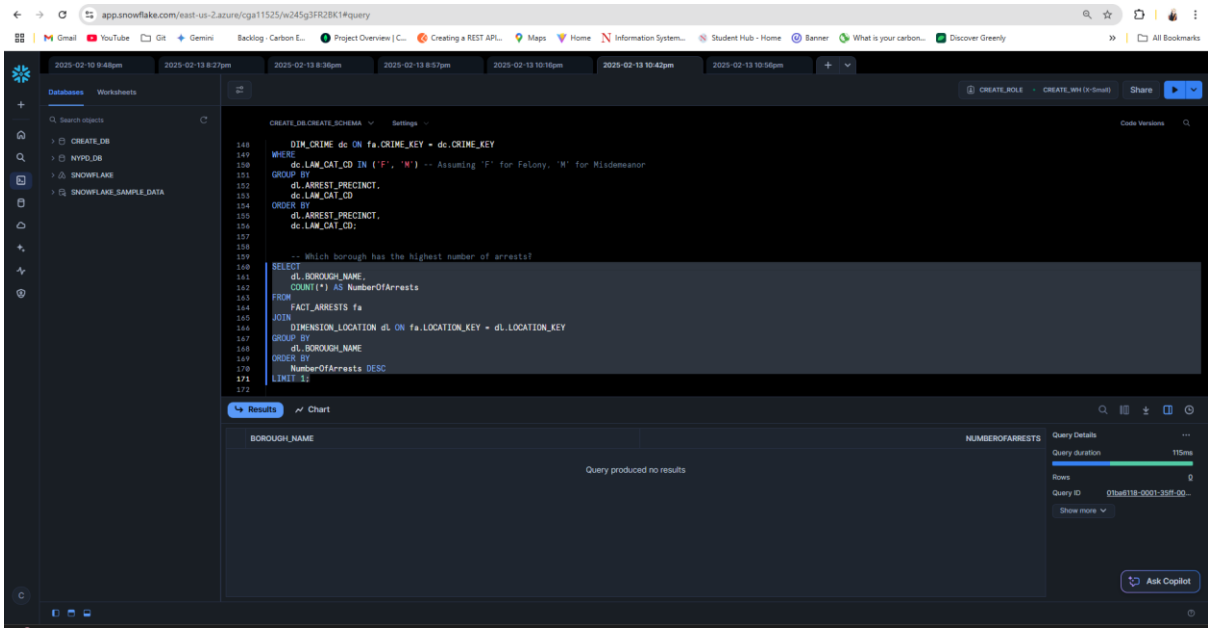
```
-- Are there specific precincts with higher felony arrests compared to misdemeanors?
SELECT
  dl.ARREST_PRECINCT,
  dc.LAW_CAT_CD,
  COUNT(*) AS NumberOfArrests
FROM
  FACT_ARRESTS fa
JOIN
  DIMENSION_LOCATION dl ON fa.LOCATION_KEY = dl.LOCATION_KEY
JOIN
  DIM_CRIME dc ON fa.CRIME_KEY = dc.CRIME_KEY
WHERE
  dc.LAW_CAT_CD IN ('F', 'M') -- Assuming 'F' for Felony, 'M' for Misdemeanor
GROUP BY
  dl.ARREST_PRECINCT,
  dc.LAW_CAT_CD
ORDER BY
```

```
dl.ARREST_PRECINCT,
dc.LAW_CAT_CD;
```



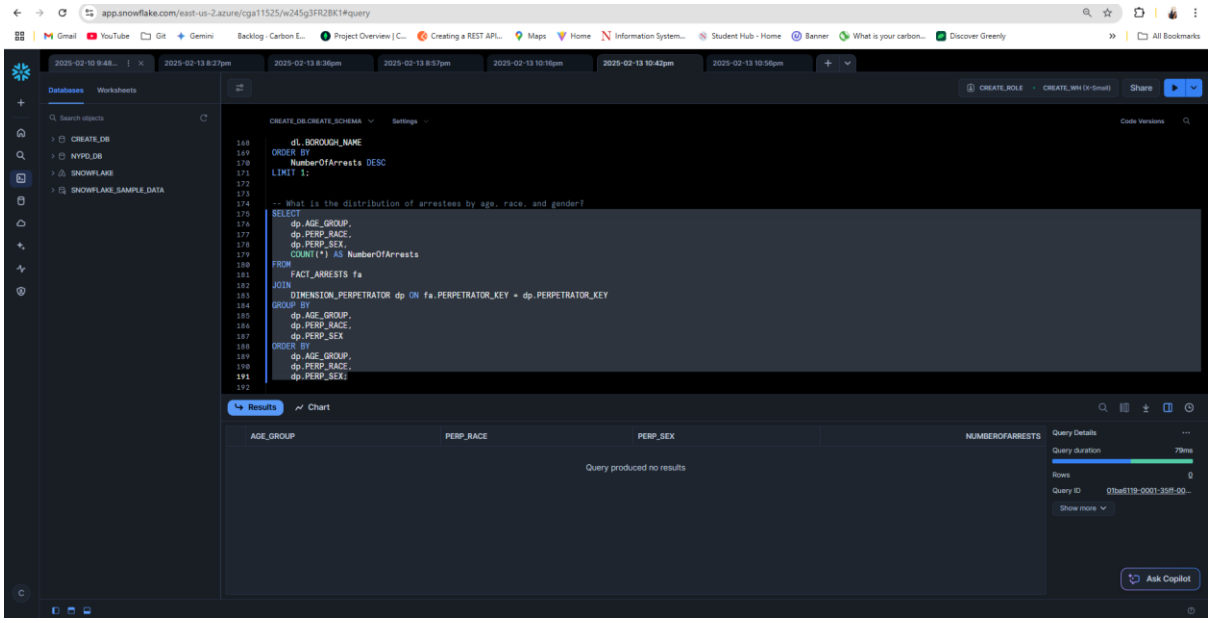
6. *Which borough has the highest number of arrests?*

```
-- Which borough has the highest number of arrests?
SELECT
  dl.BOROUGH_NAME,
  COUNT(*) AS NumberOfArrests
FROM
  FACT_ARRESTS fa
JOIN
  DIMENSION_LOCATION dl ON fa.LOCATION_KEY = dl.LOCATION_KEY
GROUP BY
  dl.BOROUGH_NAME
ORDER BY
  NumberOfArrests DESC
LIMIT 1;
```



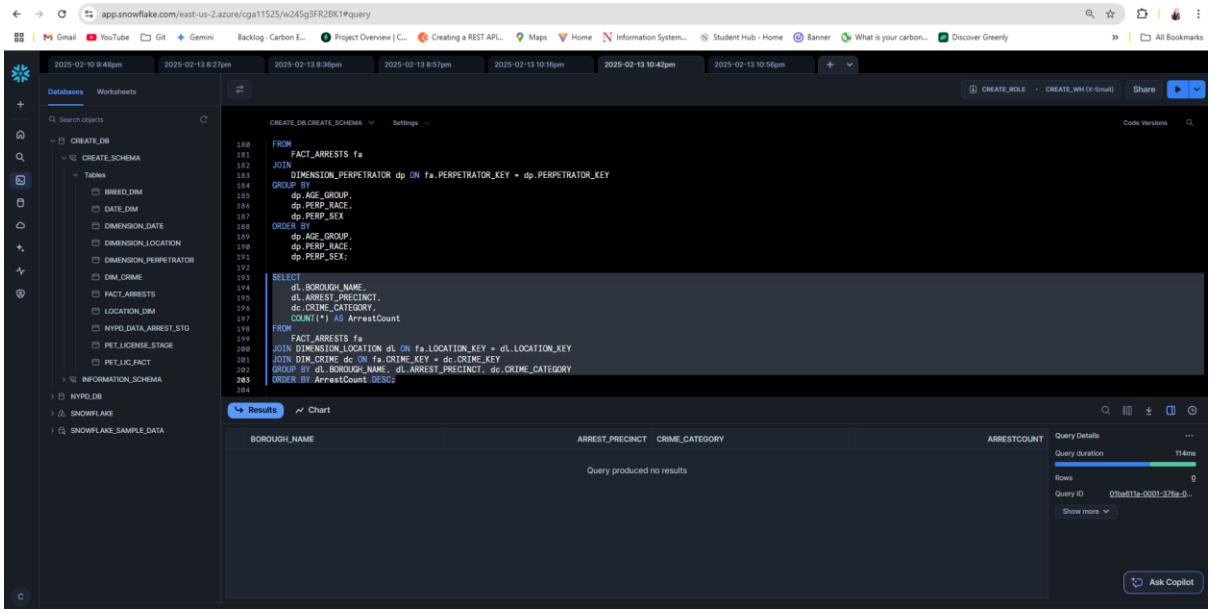
7. What is the distribution of arrestees by age, race, and gender?

```
SELECT
  dp.AGE_GROUP,
  dp.PERP_RACE,
  dp.PERP_SEX,
  COUNT(*) AS NumberOfArrests
FROM
  FACT_ARRESTS fa
JOIN
  DIMENSION_PERPETRATOR dp ON fa.PERPETRATOR_KEY =
dp.PERPETRATOR_KEY
GROUP BY
  dp.AGE_GROUP,
  dp.PERP_RACE,
  dp.PERP_SEX
ORDER BY
  dp.AGE_GROUP,
  dp.PERP_RACE,
  dp.PERP_SEX;
```



8. Predict high-crime areas based on past arrest data

```
SELECT
    dl.BOROUGH_NAME,
    dl.ARREST_PRECINCT,
    dc.CRIME_CATEGORY,
    COUNT(*) AS ArrestCount
FROM
    FACT_ARRESTS fa
JOIN DIMENSION_LOCATION dl ON fa.LOCATION_KEY = dl.LOCATION_KEY
JOIN DIM_CRIME dc ON fa.CRIME_KEY = dc.CRIME_KEY
GROUP BY dl.BOROUGH_NAME, dl.ARREST_PRECINCT, dc.CRIME_CATEGORY
ORDER BY ArrestCount DESC;
```


app.snowflake.com/east-us-2.azure/cga11525/w245g3FR2BK1#query
Databases
Worksheets
CREATE_DB
CREATE_SCHEMA
Tables
BREED_DIM
DATE_DIM
DIMENSION_DATE
DIMENSION_LOCATION
DIMENSION_PERPETRATOR
DIM_CRIME
FACT_ARRESTS
LOCATION_DIM
NYPD_DATA_ARREST_STG
PET_LICENSE_STAGE
PET_LIC_FACT
INFORMATION_SCHEMA
NYPD_DB
SNOWFLAKE
SNOWFLAKE_SAMPLE_DATA
CREATE_DB.CREATE_SCHEMA
Settings
Code Versions
CREATE_ROLE
CREATE_WH (X-Small)
Share
FROM
FACT_ARRESTS fa
JOIN
DIMENSION_PERPETRATOR dp ON fa.PERPETRATOR_KEY = dp.PERPETRATOR_KEY
GROUP BY
dp.AGE_GROUP,
dp.PERP_RACE,
dp.PERP_SEX
ORDER BY
dp.AGE_GROUP,
dp.PERP_RACE,
dp.PERP_SEX;
SELECT
dl.BOROUGH_NAME,
dl.ARREST_PRECINCT,
dc.CRIME_CATEGORY,
COUNT(*) AS ArrestCount
FROM
FACT_ARRESTS fa
JOIN DIMENSION_LOCATION dl ON fa.LOCATION_KEY = dl.LOCATION_KEY
JOIN DIM_CRIME dc ON fa.CRIME_KEY = dc.CRIME_KEY
GROUP BY dl.BOROUGH_NAME, dl.ARREST_PRECINCT, dc.CRIME_CATEGORY
ORDER BY ArrestCount DESC;
Results
Chart
BOROUGH_NAME
ARREST_PRECINCT
CRIME_CATEGORY
ARRESTCOUNT
Query produced no results
Query Details
Query duration
114ms
Rows
0
Query ID
Show more
Ask Copilot