**Individual Project 1: NYPD Arrest Data**

In this project, Alteryx and Azure Data Factory (ADF) will be used to create a Dimensional Model for NYPD arrest data analysis and to develop a Data Staging Pipeline. Supporting analytical inquiries about NYC crime trends is the main objective to assist the public and city officials in comprehending arrest trends.

**-Business Requirements:**

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

**Relevant** - The DIM\_DATE table provides DAY\_NUM, WEEK\_NUM, MONTH\_NUM, QTR\_NUM, and YEAR\_NUM for time-based analysis.

QUERY:

SELECT YEAR\_NUM, MONTH\_NUM, WEEK\_NUM, DAY\_NUM, COUNT(\*) AS total\_arrests

FROM CRIME\_FACT CF

JOIN DIM\_DATE DD ON CF.DATE\_SK = DD.DATE\_SK

GROUP BY YEAR\_NUM, MONTH\_NUM, WEEK\_NUM, DAY\_NUM

ORDER BY YEAR\_NUM DESC, MONTH\_NUM DESC, WEEK\_NUM DESC, DAY\_NUM DESC;

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

**Relevant** - DIM\_DATE allows us to identify the most common days and months.

QUERY:

SELECT MONTH\_NUM, COUNT(\*) AS total\_arrests

FROM CRIME\_FACT CF

JOIN DIM\_DATE DD ON CF.DATE\_SK = DD.DATE\_SK

GROUP BY MONTH\_NUM

ORDER BY total\_arrests DESC

LIMIT 5;

To analyze peak days:

SELECT DAY\_NUM, COUNT(\*) AS total\_arrests FROM CRIME\_FACT CF JOIN DIM\_DATE DD ON CF.DATE\_SK = DD.DATE\_SK GROUP BY DAY\_NUM ORDER BY total\_arrests DESC LIMIT 5;

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

**Relevant** - The DIM\_CRIMES table contains PD\_DESC (precinct description) and OFFENSE\_CAT (offense category).

QUERY:

SELECT PD\_DESC, COUNT(\*) AS total\_occurrences

FROM CRIME\_FACT CF

JOIN DIM\_CRIMES DC ON CF.CRIMES\_SK = DC.CRIMES\_SK

GROUP BY PD\_DESC

ORDER BY total\_occurrences DESC

LIMIT 5;

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

**Relevant** - The DIM\_DATE table allows us to analyze trends over time.

QUERY:

SELECT YEAR\_NUM, PD\_DESC, COUNT(\*) AS total\_arrests

FROM CRIME\_FACT CF

JOIN DIM\_CRIMES DC ON CF.CRIMES\_SK = DC.CRIMES\_SK

JOIN DIM\_DATE DD ON CF.DATE\_SK = DD.DATE\_SK

GROUP BY YEAR\_NUM, PD\_DESC

ORDER BY YEAR\_NUM DESC, total\_arrests DESC;

1. Are there specific precincts with higher felony arrests compared to misdemeanors? (Hint: A precinct is a police district within a city.)

**Relevant** - The DIM\_ARRESTS table contains ARREST\_PRECINCT, and DIM\_CRIMES has LAW\_CATEGORY (which can indicate felony vs. misdemeanor).

QUERY:

SELECT DA.ARREST\_PRECINCT,

DC.LAW\_CATEGORY,

COUNT(\*) AS total\_arrests

FROM CRIME\_FACT CF

JOIN DIM\_ARRESTS DA ON CF.ARRESTS\_SK = DA.ARRESTS\_SK

JOIN DIM\_CRIMES DC ON CF.CRIMES\_SK = DC.CRIMES\_SK

GROUP BY DA.ARREST\_PRECINCT, DC.LAW\_CATEGORY

ORDER BY DA.ARREST\_PRECINCT, total\_arrests DESC;

1. Which borough has the highest number of arrests? (Hint: A borough is a large administrative division in NYC, such as Manhattan (M), Brooklyn (K), Queens (Q), The Bronx (B), and Staten Island (S).)

**Relevant** - DIM\_ARRESTS contains ARREST\_BOROUGH.

QUERY:

SELECT ARREST\_BOROUGH, COUNT(\*) AS total\_arrests

FROM CRIME\_FACT CF

JOIN DIM\_ARRESTS DA ON CF.ARRESTS\_SK = DA.ARRESTS\_SK

GROUP BY ARREST\_BOROUGH

ORDER BY total\_arrests DESC;

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

**Relevant** - The DIM\_CRIMINAL table contains AGE, SEX, and RACE.

QUERY:

SELECT AGE, SEX, RACE, COUNT(\*) AS total\_arrests

FROM CRIME\_FACT CF

JOIN DIM\_CRIMINAL DC ON CF.CRIMINAL\_SK = DC.CRIMINAL\_SK

GROUP BY AGE, SEX, RACE

ORDER BY total\_arrests DESC;

1. Can we predict high-crime areas based on past arrest data?

**Partially Relevant** - DIM\_LOCATION contains LATITUDE, LONGITUDE, and GEOREFERENCE, which help identify crime hotspots. However, this dataset does not include crime severity, time of day, or socioeconomic factors necessary for a full predictive analysis.

SELECT DL.LATITUDE, DL.LONGITUDE, COUNT(\*) AS total\_arrests

FROM CRIME\_FACT CF

JOIN DIM\_LOCATION DL ON CF.LOCATION\_SK = DL.LOCATION\_SK

GROUP BY DL.LATITUDE, DL.LONGITUDE

ORDER BY total\_arrests DESC

LIMIT 10;

-Identifying the Grain:

Each record in the dataset represents a single arrest event, which is the grain of our data.

Dimension Tables:

* DIM\_DATE – Stores date-related information (day, month, quarter, year).
* DIM\_LOCATION – Stores geographic data (latitude, longitude, borough, precinct).
* DIM\_CRIMES – Stores crime-related details (crime type, law code, offense category).
* DIM\_ARRESTS – Stores arrest-related attributes (arrest ID, precinct, borough, jurisdiction).
* DIM\_CRIMINAL – Stores information about arrestees (age, race, gender).

**-**Data Profiling Using Alteryx:

The NYPD Arrest dataset was profiled using Alteryx Designer, revealing several data quality issues:

* Missing Values:
  + PD\_CD: 8 records are null
  + KY\_CD: 32 records are null
  + LAW\_CAT\_CD: 1390 records are null
  + LATITUDE: 4 records are null
  + LONGITUDE: 4 records are null
  + NEW GEOREFERENCED COLUMN: 4 records are null
* Data Formatting Issues:
  + Date format is MMDDYYYY, but should be converted to YYYYMMDD.
  + Data types need standardization for PD\_CD, KY\_CD, ARREST\_PRECINCT, JURISDICTION\_CODE, LATITUDE, and LONGITUDE.

Data Cleansing Approach:

* Replace null values with appropriate placeholders or remove invalid records.
* Convert date format for consistency.
* Standardize data types to maintain integrity in Snowflake.

Transformations:

* NULLCLEAR: Handle missing values.
* DateFormat: Convert date format to YYYYMMDD.
* Standardization: Ensure data types match schema requirements.

-ER/STUDIO:

A computer screen shot of a computer

AI-generated content may be incorrect.

A computer screen shot of a computer

AI-generated content may be incorrect.

-SQL SCRIPT:

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

-ALTERYX:

A screenshot of a computer

AI-generated content may be incorrect.

-DATAFLOW:

A screenshot of a computer

AI-generated content may be incorrect.

-ADF PIPELINE:

A screenshot of a computer

AI-generated content may be incorrect.

-DATA POPULATED IN SNOWFLAKE DATABASE:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

-ROW COUNTS USING SQL QUERY:

A computer screen shot of a program

AI-generated content may be incorrect.

-GITHUB:

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

GitHub Link: <https://github.com/varana2001/DAMG7370SPRING25>

DataFlow Link: <https://github.com/varana2001/DAMG7370SPRING25/tree/main/dataflow>

Pipeline Link: <https://github.com/varana2001/DAMG7370SPRING25/tree/main/pipeline>