Individual Project 1: NYPD Arrest Data (Part 1)

Bhagyashri Avinash Pagar **NUID: 002310690**

Introduction

The NYPD Arrest Dataset provides a comprehensive record of all arrests made by the New York City Police Department (NYPD) during the current year. Managed by the NYC Open Data Team at the Office of Technology and Innovation (OTI), this dataset is part of the city's commitment to making government data accessible to all.

The dataset, manually extracted and reviewed quarterly by the Office of Management Analysis and Planning, includes key details such as the type of crime, location, time of enforcement, and suspect demographics. It serves as a valuable resource for the public, policymakers, and researchers to explore and analyze patterns in police enforcement activity across NYC.

NYC Open Data works closely with city agencies and the civic tech community to enhance data accessibility and transparency. Each agency has an Open Data Coordinator responsible for managing and structuring public datasets. For further insights, users are encouraged to refer to the dataset footnotes for additional details.

Overview

Basic Information

- Total Number of records/rows/lines: 260,503
- Total Number of columns/fields: 19 columns

SR. NO	Column Name	Description	Input Data Type	Expected Schema Data type
1	ARREST_KEY	Randomly generated persistent ID for each arrest	V_WString	V_String
2	ARREST_DATE	Exact date of arrest for the reported event	V_WString	Date
3	PD_CD	Three-digit internal classification code (more granular than Key Code)	V_WString	Int16
4	PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)	V_WString	V_WString
5	KY_CD	Three-digit internal classification code (more general category than PD code)	V_WString	V_WString

6	OFNS_DESC	Description of internal classification corresponding with KY code (more general category than PD description)	V_WString	V_WString
7	LAW_CODE	Law code charges corresponding to the NYS Penal Law, VTL and other various local laws	V_WString	V_WString
8	LAW_CAT_CD	Level of offense: felony, misdemeanor, violation	V_WString	V_WString
9	ARREST_BORO	Borough of arrest. B(Bronx), S(Staten Island), K(Brooklyn), M(Manhattan), Q(Queens)	V_WString	V_WString
10	ARREST_PRECINCT	Precinct where the arrest occurred	V_WString	Int16
11	JURISDICTION_CODE	Jurisdiction responsible for arrest. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non-NYPD jurisdictions	V_WString	Int16
12	AGE_GROUP	Perpetrator’s age within a category	V_WString	V_WString
13	PERP_SEX	Perpetrator’s sex description	V_WString	V_String
14	PERP_RACE	Perpetrator’s race description	V_WString	V_String
15	X_COORD_CD	Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)	V_WString	Double
16	Y_COORD_CD	Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)	V_WString	Double
17	LATITUDE	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	V_WString	Double
18	LONGITUDE	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	V_WString	Double
19	NEW_GEOREFERENCED_COLUMN	geocoded column	V_WString	Spatial Object

Data Quality Assessment

Issues	Column Name	Details	Action Plan
Missing Values	LAW_CAT_CD	It has 0.53% missing Values. There are 1390 values which are missing or null	Replace the missing values with unknown
Inconsistent Format	LATITUDE, LONGITUDE, NEW_GEOREFERENCED_COLUMN	These columns are not in the uppercase, and all other columns are in uppercase	Rename these columns in uppercase
	ARREST_DATE	The date is in the format mm/dd/yyyy We need to convert that date into yyyy-mm-dd format	Use DateTime tool to convert the data into the right format using Alteryx
	LAW_CAT_CD	This column should only have three values for Level of offense: felony(F), misdemeanor(M), violation(V) but, it also contains some other values such as 9 and I	Replace the invalid values from the column with unknown.
Null Values	PD_DESC	It contains 8 null values.	Replace those null values with unknown
	KY_CD	This column contains 0.015 null values	Replace those values with -1
	OFNS_DESC	It contains 32 null values	Replace those values with unknown

	LAW_CODE	Contains 8 null values	Replace those null values with -1
	LAW_CODE_CD	Contains 8 null values	Replace the null values with unknown
	PD_CD	This Column contains 8 null values	Replace the null values with -1
Mismatched Data Type	All the columns (19 columns of given dataset)	Almost all the columns have different data types when we load the tsv file in Alteryx	Assign the appropriate data type to all the columns based on the description of the data

Summary Statistics

Numerical Features

Columns such as PD_CD, KY_CD, ARREST_PRECINCT, JURIDICTION_CODE, X_COORD, Y_COORD, LATITUDE, LOGITUDE these columns contain numerical values but there is no significance of deriving Min, Max, Mean, Median, Std deviation for these columns.

Categorical Features

Columns such as PD_DESC, OFNS_DESC, ARREST_BORO, PERP_SEX, PERP_RACE are some of the categorical features of the dataset.

Data Transformation

- Remove the null and missing values present in the columns.
- Standardize the date format.
- Add new columns such as DI_JOB_ID and DI_LOAD_ID.
- Standardize all the column names into a consistent format.
- Assign appropriate data types to each of the columns.

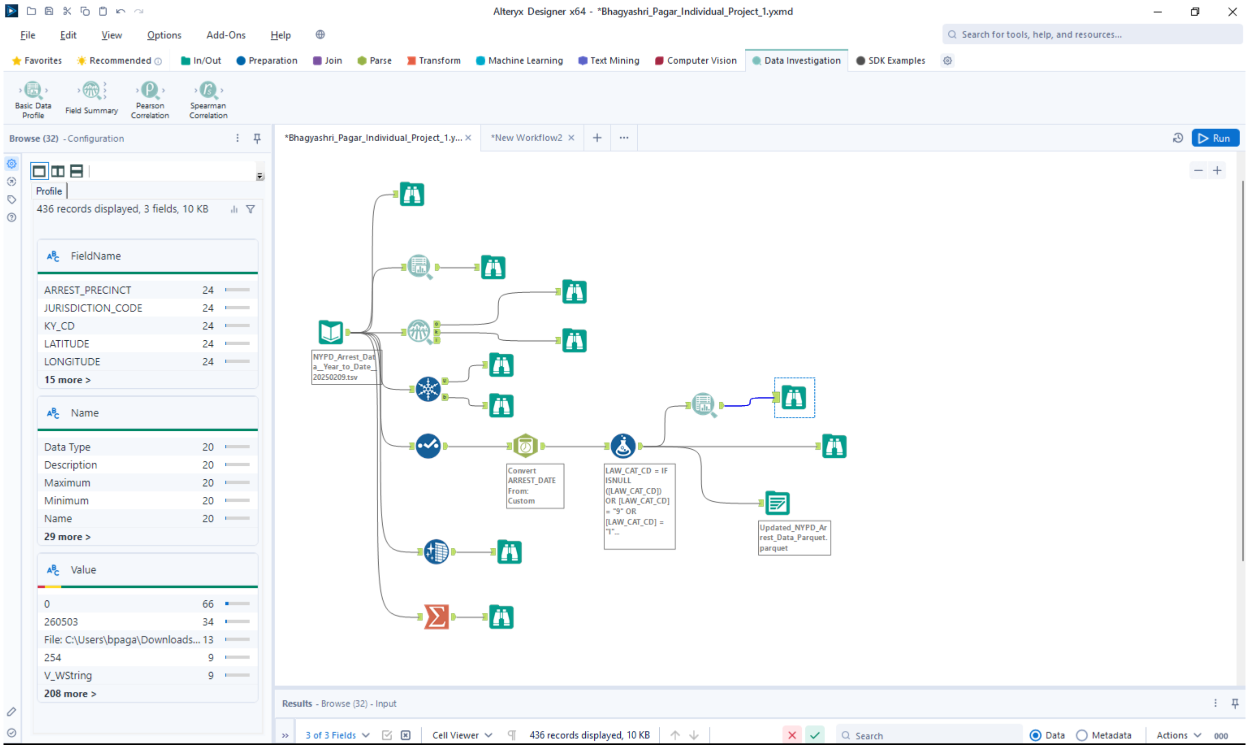
Insights and Observations

- Based on data profiling observations, there are some missing and null values in some columns. Although the percentage of missing values is very less, we still need to handle it because sometimes it may give wrong observations.

Conclusion and Recommendations

- Make the formatting of all the columns with text fields consistent.
- Remove all the null and missing values from the data.
- Convert all time stamps to a uniform format for accurate time-based analysis.

Alteryx Workflow Used for Data Profiling


Alteryx Designer x64 - *Bhagyashri_Pagar_Individual_Project_1.yxmd
File
Edit
View
Options
Add-Ons
Help
Search for tools, help, and resources...
Favorites
Recommended
In/Out
Preparation
Join
Parse
Transform
Machine Learning
Text Mining
Computer Vision
Data Investigation
SDK Examples
Basic Data Profile
Field Summary
Pearson Correlation
Spearman Correlation
Browse (32) - Configuration
*Bhagyashri_Pagar_Individual_Project_1.y...
*New Workflow2
Run
Profile
NYPD_Arrest_Data__Year_to_Date_20250209.tsv
Convert ARREST_DATE From: Custom
LAW_CAT_CD = IF ISNULL ([LAW_CAT_CD]) OR [LAW_CAT_CD] = "9" OR [LAW_CAT_CD] = "I"...
Updated_NYPD_Arrest_Data_Parquet.parquet
Results - Browse (32) - Input
3 of 3 Fields
Cell Viewer
Search
Data
Metadata
Actions

436 records displayed, 3 fields, 10 KB

FieldName	
ARREST_PRECINCT	24
JURISDICTION_CODE	24
KY_CD	24
LATITUDE	24
LONGITUDE	24
15 more >	

Name	
Data Type	20
Description	20
Maximum	20
Minimum	20
Name	20
29 more >	

Value	
0	66
260503	34
File: C:\Users\bpaga\Downloads...	13
254	9
V_WString	9
208 more >	

436 records displayed, 10 KB

Github Link:

https://github.com/Pagar-Bhagyashri/DAMG7370/tree/Individual_Project_NYPD_Arrest_Data