NYC Council Data Scientist Exercise

Objective:

The goal of this exercise is to assess your programming skills and problem-solving approach. There isn't a single correct answer or method. We're interested in clear and well-justified solutions.

Hypothetical Scenario:

A staffer from the Council's Public Safety Committee is preparing a briefing paper and questions for a hearing on NYPD arrests. They are interested in knowing if the arrest rate has decreased, understanding overall trends in arrests, and what we can say about police enforcement in the city. They have approached you, as a member of the Data team, for an analysis of arrests from 2019-2024.

About the NYPD Arrests Data Set (Historic):

The NYPD Arrests Data Set (Historic) provides incident-level data related to arrests in NYC, including the type of crime, location, time of enforcement, and suspect demographics, as noted on the Open Data Portal. The data and metadata can be accessed below.

- NYPD Arrests Data (Historic)
- NYPD Arrest Data (Year to Date)
- NYPD Arrest Incident Level Data Footnotes

Deliverables:

Validate, clean, and analyze the NYPD data set. Answer the following questions in a report of 2-3 pages (excluding figures).

1. **Arrest Rate Trend**: Has the arrest rate decreased from 2019 to 2024? Describe the trend and justify any statistical tests used.
2. **Top 5 Arrests**: Identify the top 5 most frequent arrests in the 'pd_desc' column from 2019-2024. Describe and compare the trends over time.
3. **Crime Comparison**: Is there more crime in Precinct 19 (Upper East Side) compared to Precinct 73 (Brownsville), using arrests as a sample? Describe the trend and variability, and justify any statistical tests used.
4. **Predictive Model**: Propose a model to predict crime for better resource allocation by the NYPD. Discuss potential challenges, the choice of independent and dependent variables, and how you would evaluate the model.

Clearly explain and justify any assumptions, statistical tests, plots, and models you use. Feel free to use external publicly available datasets and include any exploratory data analysis that adds value to the analysis. You may use any programming language you are comfortable with, preferably R or Python.

Include all your code, comments, and the report in a new private GitHub repository under your username. Once completed, provide read access to the repository by inviting @AlaaMoussawi, @romartinez-nycc, @melissanunezcouncil, and @nmontalbanocouncil.