NYC Council Data Scientist Exercise

Objective:

The goal of this exercise to get an understanding of your programming skills and the steps you take to solve problems. There is no one single answer or way to approach the exercise. We are looking for coherent and justified solutions.

Hypothetical Scenario:

A staffer from the Council's Public Safety Committee is preparing a public briefing paper and questions for a hearing on NYPD arrests. The staffer wants to know if the arrest rate has decreased. They also want to know about some overall trends in arrests and what we can say about police enforcement in the city. They have come to you as a member of the Data team asking for an analysis of arrests from 2015-2018.

About the NYPD Arrests Data Set (Historic):

The NYPD Arrests Data Set (Historic) includes incident level data related to arrests. As stated in the Open Data Portal, "each record represents an arrest effected in NYC by the NYPD and includes information about the type of crime, the location and time of enforcement. In addition, information related to suspect demographics is also included".

The data & metadata can be accessed below.
https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u
NYPD Arrest Incident Level Data Footnotes

Deliverables:

Please validate, clean, and analyze the NYPD data set and respond to the following questions from the Council staffer in a report not exceeding 2-3 pages (excluding figures):

- **Has the arrest rate been decreasing from 2015-2018?** Describe the trend and defend any statistical tests used to support this conclusion.
- **What are the top 5 most frequent arrests as described in the column 'pd_desc' in 2018?** Compare & describe the overall trends of these arrests across time.
- **If we think of arrests as a sample of total crime, is there more crime in precinct 19 (Upper East Side) than precinct 73 (Brownsville)?** Describe the trend, variability and justify any statistical tests used to support this conclusion.
- **What model would you build to predict crime to better allocate NYPD resources?** What challenges do you foresee? What variables would be included? How would you evaluate the model? Discuss in no more than 150 words**.**

Include any exploratory data analysis that you feel adds to the analysis.
Clarify and justify any assumptions and statistical tests, plots, and models used.
Feel free to use external publicly available datasets if needed.

Please use any programming language you are comfortable with (preferably R or Python).

Send us your code, comments, and report via the **Data Exercise** form included in the email. In the form, you can either upload your results or provide a GitHub repository link.