Assignment 1: Exploratory Visual Analysis

Please present your work on Tuesday March 29 during the class

In groups of 1–3, download NYPD Stop and Frisk dataset and perform exploratory analysis using any visualization tool/programming language/library. Then prepare a PDF or slides:

Task 1: You should seek to gain an overview of the shape & structure of your dataset. What variables does the dataset contain? How are they distributed? Are there any notable data quality issues? Are there any surprising relationships among the variables?

Task 2: Investigate hypotheses (questions) and develop preliminary insights. Then include a set of 6 or more visualizations that illustrate your findings and a write-up of your process and what you learned. For each question, create a visualization that might provide a useful answer.

Task 3: Answer the following questions. Use visualization techniques to make the answers visible.

- Who is most affected by Stop and Frisk, and in what capacity are they affected?

- When does Stop and Frisk happen the most?

- Where in New York City do the stops occur?

- Why do officers stop people?

- What are the most important factors contributing to a person getting frisked

- Use a classification model to predict whether a stop is “effective”. Define “effective”. Report AUC and ROC curve.

- Use a classification model to tell whether someone will be frisked or not? Report AUC and ROC curve.

Final Deliverable

Your final submission will be a written report, divided into three parts (tasks)

Each visualization image should be accompanied with a title and short caption (<2 sentences). Provide sufficient detail for each caption such that anyone could read your report and understand your findings. Feel free to annotate your images to draw attention to specific features of the data.

Grading Criteria

- Clear questions.
- Appropriate data quality assessment and transformations.
- Breadth of analysis, exploring multiple questions.
- Depth of analysis, with appropriate follow-up questions.
- Expressive & effective visualizations appropriate to analysis questions.
- Clearly written, understandable captions that communicate primary insights.

Data Source

- NYC Open Data: data on NYC trees, taxis, subway, citibike, 311 calls, land lot use, etc.
- Background reading: NY Civil Liberty Union Stop and Frisk statistics: https://www.nyclu.org/en/stop-and-frisk-data
- Dataset source: NYPD Stop and Frisk Data: https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

Additional Tools

Your dataset almost certainly will require reformatting, restructuring, or cleaning before visualization. Here are some tools for data preparation:

- Graphical Tools
    - Tableau includes basic functionality for data import, transformation & blending.
    - R with ggplot2 library
    - Python Jupyter notebooks with libraries eg. Altair or Matplotlib
    - Trifacta Wrangler interactive tool for data transformation & visual profiling.
    - OpenRefine free, open source tool for working with messy data.
- Programming Tools
    - JavaScript data utilities or Datalib JS library via Vega.
    - Pandas data table and manipulation utilites for Python.
    - dplyr an R library for data manipulation.
        - Or, the programming language and tools of your choice.