**Data Mining (Fall 2024)**

**Homework 1**

**Data Wrangling with Baltimore City 911 Report Data**

**Due before class on 9/30/2024**

**The dataset can be downloaded from this** [**folder**](https://drive.google.com/drive/folders/1-1xW8r62LZnPX-Lam5gINGZVrQ4bc4q5?usp=sharing) **(Baltimore911.csv)**

You were just hired as a data analyst for a non-profit organization in Baltimore called “Safe Neighbourhood” which uses data-driven insights to support initiatives to improve public safety. On the first day of your job, you attend a meeting with Chief Safety Officer Gary who is a retired police officer. He brings your attention to an interesting publicly available dataset curating the 911 call reports of the last decade. He asks whether you could help him to turn those data into useful insights to (1) help the police department decide on where and when to deploy resources; (2) provide safety advisory to residents/visitors on the likelihood of crime and types of crime at given time and given location. As the first step to solving this problem, you decide to do an exploratory data analysis to identify interesting patterns to discuss with Gary in the next meeting.

After coming back from the meeting, you explored the data that Gary shared with you and created the following to-do list:

1. Create a dataset profile table that gives an overview of the dataset. You may refer to profile tools such as <https://oralytics.com/2022/04/04/python-data-profiling-libraries/>

At a minimum, you will need to gather the following information from the dataset via writing Python scripts where appropriate.

* Total number of call instances/rows
* Total number of call features/columns
* For each column:
  + Are they numeric/real/continuous or symbolic/discrete types of attributes or belong to temporal or spatial categories?
  + If it is a numeric attribute, what are the min, max, mean, median, and standard deviation of the values?
  + If it is a discrete attribute, what is the total number of unique values? What are the three attribute values with the largest count?
  + What is the level of missingness (% of rows with missing values)?

Comment on anything surprising/strange/unique you noted from the data profile.

1. Generate a series of plots to describe the temporal pattern (year-to-year, monthly, and day-of-week) of the **overall crime** incidence aggregating from all geo-locations;
2. Generate a plot describing the distribution of crime type aggregating from **all geo-locations and all time periods;**
3. Generate a series of plots to illustrate how crime type distributions might vary from year-to-year, month-to-month, or by day-of-week aggregating across all locations.
4. Design a dashboard that allows users like Gary to explore the spatial and temporal patterns of crime. You may get inspiration from tasks 2-4, but feel free to add insights. **Bonus: Create a real dynamic dashboard that allows users to change parameters such as location and time period.**

**Deliverable:**

* A Google slide deck ready to be presented to Gary summarizing results from the tasks above. Please label clearly on the slide which task you are tackling. There are no lower/upper limits on the number of slides. Always keep the message concise and effective, keeping your audience in mind. In this case, it is your colleague like Gary who might not be trained in data analytics/data science.
* Please make sure the plots are professionally formatted (e.g. with clear labeling of the axis, legible font size, and reasonable color scheme, etc.). In each slide, you will add a textbox telling a story to the audience about what is notable/interesting/surprising, etc. Image yourself as a tour guide of the data.

**Software and Tools**

Python is highly recommended. You may also use Tableau.

**GPT Policy**

You may use GPT for this task. If you use GPT, please include a separate document briefly summarizing (1) what you used GPT for, (2) your reflection/experience using GPT, and (3) some example prompts that you find useful.

**Submission:**

* Please create a GitHub account and upload codes and *optionally other files necessary to create your portfolio. You may create a GitHub page to present your project, but this is optional for this homework assignment.*
* For submission on Blackboard
  + A link to your Google slides (Please change the slide's permission to editable so that we can comment/provide feedback on your slides, see instructions below on how to share correctly

| Select the file you want to share.  Click Share or Share Share.  Under “General access” click the Down arrow Down.  Choose Anyone with the link.  To decide what role people will have, select Editor.  Click Copy link.  Click Done.  Paste the link in an email or any place you want to share it. |
| --- |

* A link to your GitHub account
* GPT statement, if applicable