EAS 503 - Project Requirements

The ideal goal of the project is to create a data science pipeline that allows users to understand a real world problem or phenomenon through an interactive dashboard that feeds off a realistic data source. The focus should be on analyzing real life data to make some meaningful interpretations and conclusions.

For instance, the Chicago crime data (used in class) can be used to identify the high crime zones within the city of Chicago. Each group has the liberty to choose a suitable problem and corresponding datasets for the project. A list of datasets that can be used are provided.

1 Project Logistics

Students need to to form groups of 3-4 members by October 18^{th}.

Project consists of four main submissions:

1.1 Preliminary report

Each group will submit a preliminary report (max 1 page) that includes introduction, data description, proposed analysis, analysis methods that will be used for the project, milestones, and references. No more than 2 groups will be allowed to use the same dataset.

1.2 Final Presentation

A 12-minute presentation about the goals set, the dataset used, the analysis methods used, and the results achieved. The presentation must be in the form of a Jupyter Notebook dashboard that contains the final results and interpretations made.

1.3 Project Final Report

A written report that includes including title, abstract, introduction, model, data, analysis and results, conclusion, future research directions, reference, and appendix.The final report without including the figures should be 1-2 pages.

1.4 Project Code

All implementations will have to be uploaded (via Github) to a repository (announced later). This includes the main notebook, any additional codes, SQL data schema, data (if not too large), etc.

Note: The report must be submitted in the form of a pdf file (12 pt. font, single spacing).

2 Deadlines

October 18	Group Member Details (Form groups on `UBLearns`)
November 8	Project Description
November 29	Preliminary Report
December 15	Final Presentation
December 20	Code and Report Due

3 Project Description

To develop a project, start with two core questions in mind:

1. What is the real life problem you want to tackle? This could be something like - understanding crime in Buffalo, or, *predicting traffic issues in the city of Chicago*, etc. While the above examples are all within the urban domain, you are free to choose problems from other sectors as well.

2. What is the data needed to develop a useful interface to handle the above problem? If you plan to develop something within the urban domain, you can make use of numerous open data repositories that many large cities (including Buffalo) have created. See below for a list of these repositories.

3.1 Expectations

From the programming perspective, the project should have following components:

- **Data in a relational database.** You will need to submit the database schema, as well as data ingestion scripts.

- **Python notebook as the primary implementation.** This should have following components:

  - *A data import component.* This component should pull the data from the database into the Python environment. Note that you will need to demonstrate that you are manipulating data in the database itself (through embedded SQL) and not just pulling all the data into the notebook (no `SELECT * FROM table`). Bulk importing will result in loss of points.

  - *A data analysis component.* This is where you will perform various types of analyses (including basic histogramming, plotting, etc., and optionally advanced predictive analytics, clustering, etc.). You are allowed to use any Python package discussed in class for this.

  - *A visualization component.* This is where you will utilized the visualization capabilities of the MatplotLib library to show the output of your analyses. Choose the visualization outputs wisely! They should support your starting question and should allow you to draw conclusions. In some cases, it might make sense to put data on a map. In other cases, an interactive portal, where users can "play" with the data would be useful. We leave that to your design choices.

3.2 Grading

You will be graded out of a total 100 points. While an exact rubric will be circulated towards the later part of the course, the general scheme would be to assign points based on:

1. Submission of all required components
2. Presence of all required programming aspects in the submitted code
3. Real world relevance of the problem
4. Presentation skills

We reserve 10 extra points for creativity and use of advanced technical components.

4 Potential Datasets

Here is a short list of data portals offered by various cities in United States. You are free to explore other portals. If you find any interesting ones, please share them on Piazza.

Buffalo	https://data.buffalony.gov/
Chicago	https://data.cityofchicago.org/
New York State	https://data.ny.gov/
Boston	https://data.boston.gov/
San Francisco	https://datasf.org/opendata/
New York City	https://opendata.cityofnewyork.us/
Los Angeles	https://data.lacity.org/
Washington DC	http://opendata.dc.gov/
Baltimore	https://data.baltimorecity.gov/
Minneapolis	http://opendata.minneapolismn.gov/
Kaggle	https://www.kaggle.com/