SDS 3786 Lab 1: Les aspects non-techniques de l'analyse des données.

Professor: Patrick Boily
Anthony Le, John Bute, Kelly Gao

Introduction of Our team (John)

John

- **Areas of Interest:**
  - Statistical Methods
  - Machine Learning
  - Data Engineering
  - Big Data Technologies

Anthony

- **Areas of Interest:**
  - Data Translation (Explaining Data in Contexts)
  - Data Visualization
  - Data Cleaning
  - Machine Learning

Kelly

- **Areas of Interest:**
  - Deep Learning
  - Big Data
  - Artificial Intelligence (AI)

Complementary Strengths and Identified Gaps

- **Complementary Skills:**
  - John and Kelly bring a strong focus on data engineering, machine learning, and big data, with Kelly adding expertise in deep learning and AI.

    - Anthony complements the technical strengths of the team with his skills in data translation and communication, providing context to data insights, which is essential for report writing and data presentation.
- **Identified Gaps:**
    - **Programming and Technical Depth**: John has weaknesses in programming and deep learning, which are partially covered by Kelly. However, we have a notable gap in NLP and advanced statistical methods across the team.
    - **Mathematical Algorithms and Data Science Fundamentals**: Anthony lacks experience with mathematical algorithms and data science-specific technical skills, which limits the depth of analytical capabilities in the team.
    - **Ethics:** Since everyone is not familiar with the industry and its ethical standards regarding the use of data and AI, there is a notable gap that needs to be addressed promptly.

By addressing these gaps, we can improve its overall analytical capability and ensure a well-rounded approach to data-driven problem-solving.

Ethical Declaration (Kelly)

We will pursue AI/ML work with prioritization of privacy and data protection. We will act upon our projects with fairness and without bias, and we will pursue objectives and results that enhance human capabilities, not detract from them. We recognize the profound impact that our work can have on individuals and our peers.. We pledge to prioritize privacy, data protection, fairness, and human-centered design in all our projects.

1. Privacy and data protection. All information will be from publically available datasets with publicly available information that does not infringe on personal rights, while complying with common privacy concerns. This principle is important to us as a team as the increasing availability of personal data obtained online that is fed into AI and ML applications raises significant concerns around user privacy and data misuse. We are committed to safeguarding the privacy of individuals by ensuring that data used in our models is collected, stored, and processed with strict adherence to data protection regulations and best practices. Respecting user privacy is fundamental to integrity of our work.
2. Fairness and bias. In our course, the datasets we use will come from public sources, surveys, or personal data, which may carry inherent demographic biases- gender, racial, etc. Although inherent bias may occur, we will take care to ensure that our models don't unfairly favour certain groups or amplify any inherent biases. Even though some biases in data may reflect real-world patterns, fairness and non-discrimination are essential to prevent further harm, promote equal opportunity, and reduce systemic inequalities. This principle is vital to our contribution to a more equitable and just society.
3. Human-centered design. We strongly believe that AI is best used as a tool to assist and optimize tasks. AI should be designed to augment human capabilities, not replace them. In our ideal world, scientists and engineers remain in control of technology and systems are developed to enhance human decision-making at most; we will design our projects

with the principle of maintaining human attributes such as empathy and ethical judgement.

In conclusion, we are dedicated to building AI/ML systems that align with the principles of privacy, fairness, and human-centered design. Our mission is to create technology that enhances human well-being and drives positive societal change, all while upholding the highest ethical standards.

General Structure of the Datasets (Anthony)

The general structure of the PIMENTO_PROGRAMS dataset is composed of multiple rows of seemingly duplicate information at first, with the rows identifying a geographic region, mission title, and an employee code. However, with further inspection of all of the other columns, we can observe that the dataset focuses on a variety of employees who have worked or are currently working on a mission in a specific region, and the sorts of different aspects of that mission that the employee works on during that time, like a time log. This is reinforced by the fact that the values for each column being mostly in increments of 60, 90, 120, or 240, which could be the converted hour-to-minute values for each employee. Below is a dictionary listing each of the columns and the potential definition for each one.

Definitions for PIMENTO_PROGRAMS: (Anthony)

- Other (int): Time spent working in miscellaneous tasks for the project.
- Emergency (int): Time spent on emergency-based activities (in minutes) for the project.
- Program_Mgmt (int): Time spent working around the project's Program Management.
- Liaison (int): Time spent providing liaison to the project.
- Visit_Mgmt (int): Time spent working on the program's Visit Management.
- Pol_Econ (int): Time spent working on Political Economics for the project.
- Comm_trade (int): Time spent working on commercial trading tasks for the project.
- Development (int): Time spent working on developmental tasks for the project.
- Police (int): Time spent working on police-related matters for the project. (?)
- Immigration (int): Time spent working in Immigration services for the program.
- Program_Services (int): Time spent working in or providing program services.
- Public_Comms (int): Time spent working in public communications.
- Training (int): Time spent training during work.

For the PIMENTO_CASES file, the general structure of this dataset is similar to the first file, but with different column values and names. This dataset presents the number of cases that an employee handled on a specific day. These column names denote the kinds of cases an employee may face during the day and the amount of time an employee faced to resolve those specific cases. Below is a dictionary listing each of the columns and the potential definition for each one.

Definitions for PIMENTO_CASES: (Anthony)

Note: values displayed in these columns for each row are the number of cases an employee handles and the time spent on those cases for each day.

- Other (int): number of other miscellaneous cases worked on by the employee.
- Disaster (int): Number of Disaster cases handled by an employee.
- Disaster time (int): time spent handling those disaster cases by an employee.
- Death and Death Time (int, int): number of deaths and the time spent processing them.
- Assistance Communications and ‘’ + time (int, int): number of cases in assisting communication and the time spent on those cases.
- Legal/Notary and Legal/Notary time (int, int): number of legal/notary cases handled by the employee and the amount of time spent on those cases.
- Evacuation and Evacuation time (int, int): number of Evacuation cases handled by the employee and the amount of time spent on those cases.
- Child Abduction/Custody and time (int, int): number of child abduction and custody-related cases and the time spent on them by the employee.
- Family Distress and Family Distress time (int, int): number of family distress cases and the time spent on it by employee.
- Registration and time (int, int): number of cases involving registration and the time spent on it by employee.
- Arrest and time (int, int): number of cases involving arrests and the time spent on it.
- Citizenship and time (int, int): number of citizenship cases and the time spent on it.
- Passport and time (int, int): number of passport cases and time spent on it.
- Service and time (int, int): number of service cases and time spent on it.
- Financial Assistance and time (int, int): number of financial assistance cases and time spent on it.

Questions about the dataset (Anthony)

With the dataset now defined, we have a variety of questions that we would like to answer through further investigation about the dataset itself:

- Why are certain days present in the time logs for both PIMENTO_PROGRAMS and PIMENTO_CASES, but other days are not?
- Do certain programs in PIMENTO_PROGRAMS have a correlation to the cases in PIMENTO_CASES? Would certain cases involving death or disaster increase the amount of time an employee is working on Emergency tasks for the program?
- Do the total minutes from all activities in PIMENTO_PROGRAMS and PIMENTO_CASES done from a single employee in one day equate to eight hours? Are they underworking or overworking?
- Why is the Program_Management column the most non-zero column in the dataset? Are the employees required to work on Program Management to a consistent level?

- Can the total average amount of minutes worked by the employees on a project indicate the overall state of the project? Would there be other conditions that may affect both the project's success and the hours worked by an employee?

Team Goals and Focus (John)

Our main goals for future projects are to improve our technical skills in data handling, model development, and result interpretation through report writing. We aim for these labs and our final project to equip us with the skills needed.

As a team, we will:

- Teach each other our individual strengths.
- Make a collective effort to address common weaknesses.
- Focus on building each other's strengths early to manage the complexity of advanced data science techniques later in the course.

Scheduling Plan

We will follow a structured approach to ensure timely completion of our lab work:

- **Skeleton Lab Report:** Completed for every lab about 4 days before the lab session.
- **Lab Work Completion:** 50% of the lab will be completed by the start of the lab session to allow for comfortable completion of the remaining tasks and the report by the end of class.

We will employ the SMART objectives framework to:

- Set goals for each lab.
- Define what we want to learn.
- Establish the time scope for each lab.

Project Schedule

- **September 23:** Decide on our independent project's subject.
- **September 30:** Outline the step-by-step process to complete the project.
- **October 14:** Finish the project proposal.
- **October 11:** Complete the progress report of the project.
- **December 2:** Presentation and final report completed.

Communication Strategy

- We already communicate daily, allowing us to discuss the project and provide updates regularly.
- Additional time slots will be allocated for group work on labs and the project.

- A Slack group chat has been set up for updates and questions.

Technical Focus

- Emphasis on quality control and risk management.
- Regular check-ins with our instructor, Patrick Boily, for feedback and code adjustments.
- Utilize instructor-provided resources and online tools when stuck.
- Weekly backups on a 1TB hard drive to prevent data loss due to unforeseen errors.

Submission and Presentation Strategy

- Timely submission of lab reports, at least 3 hours before the due date.
- Final project submissions and presentation preparations will be done well in advance to ensure mastery of our subject.

Role Definition

Roles are clearly defined but will rotate regularly to ensure a balanced learning environment. Team members will support each other to grow together and derive meaningful lessons from this course.