**Group Project (Spring 2023)**

**Guidelines:**

* Address empirical questions about one of the two topics, or any other topic you think is interesting.
* Python is recommended but you can use any software tool.
* A team of 3;
* Feel free to add additional data or use your own data if necessary.

1. Submit a short **write-up** before you work intensively on the analysis (March 27th 2023). Your write-up should briefly discuss your proposal, describe the dataset, and explain the potential variables you use. Less than one single-spaced page.
2. Submit a **draft report** before the end of the semester (May 1st 2023). Your draft report should include the following two parts:
   1. An executive summary: A brief description, one single-spaced page or less, describing what’s great about your analysis and your graphic.
      1. Highlight the uniqueness of your analysis and findings.
   2. A visualization of the analysis: A graphic that reflects the major steps and findings of your data analysis.
      1. It can be a static image or interactive (such as JavaScript). Interactive graphics must be self-contained and run without special software on any Windows 10 PC. Web-based graphics must run using the Chrome browser.
3. Submit a **final report** by the end of the semester (May 15th 2022).

**Your entry will be disqualified if…**

* Either of the write-up or the draft/final report is submitted after the deadline.
* You don’t specify the project you are addressing in your final report.
* Your names and IDs are not on both the submissions.
* The attachments won’t open or are in the wrong file format.

**How entries will be evaluated**

* Visualization (6 points)
  + Clarity (how well the graphic stands on its own without additional explanation)
* Analysis (10 points)
  + Novelty/creativity (originality of thought; surprising way of approaching the data)
  + Relevance (analysis relates to the problem statement)
  + Depth (sophistication of the analysis)
  + Completeness (degree to which the analysis answers the stated question)
  + Implication (ability of the graphic to aid decision making)

**Topic 1: Motor Vehicle Collisions - Crashes**

Data: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95 (please use the first 1 million rows, due to the limit of Excel)

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC. The police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage (https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/ny\_overlay\_mv-104an\_rev05\_2004.pdf). It should be noted that the data is preliminary and subject to change when the MV-104AN forms are amended based on revised crash details.For the most accurate, up to date statistics on traffic fatalities, please refer to the NYPD Motor Vehicle Collisions page (updated weekly) or Vision Zero View (updated monthly).

Questions: Which factors do you think lead to high risks of death from motor vehicle crash? Why?

**Topic 2: Data Science Survey**

Data: https://www.kaggle.com/competitions/kaggle-survey-2021/overview

This year, as in 2017, 2018, 2019, and 2020 we set out to conduct an industry-wide survey that presents a truly comprehensive view of the state of data science and machine learning. The survey was live from 09/01/2021 to 10/04/2021, and after cleaning the data we finished with 25,973 responses!

There's a lot to explore here. The results include raw numbers about who is working with data, what’s happening with machine learning in different industries, and the best ways for new data scientists to break into the field. We've published the data in as raw a format as possible without compromising anonymization, which makes it an unusual example of a survey dataset.

This year Kaggle is once again launching an annual Data Science Survey Challenge, where we will be awarding a prize pool of $30,000 to notebook authors who tell a rich story about a subset of the data science and machine learning community.

In our fifth year running this survey, we were once again awed by the global, diverse, and dynamic nature of the data science and machine learning industry. This survey data EDA provides an overview of the industry on an aggregate scale, but it also leaves us wanting to know more about the many specific communities comprised within the survey. For that reason, we’re inviting the Kaggle community to dive deep into the survey datasets and help us tell the diverse stories of data scientists from around the world.

The challenge objective: tell a data story about a subset of the data science community represented in this survey, through a combination of both narrative text and data exploration. A “story” could be defined any number of ways, and that’s deliberate. The challenge is to deeply explore (through data) the impact, priorities, or concerns of a specific group of data science and machine learning practitioners. That group can be defined in the macro (for example: anyone who does most of their coding in Python) or the micro (for example: female data science students studying machine learning in masters programs). This is an opportunity to be creative and tell the story of a community you identify with or are passionate about!