Phase I

During Phase I your goal is to find a dataset online, and formulate an interesting prediction problem around the dataset. I leave it up to you and your teammates to work on a dataset and a problem that excites you. However, you will have to “defend” your proposed project in class **10/2**. These presentations should contain **2-3 slides** and last up to **5-6 minutes**.

During the presentation you should perform the following tasks and answer the following questions:

* Clearly state the problem you are solving.
  + By accurately forecasting the delay duration of a particular flight, the goal is to help airlines, airports, and passengers better manage their schedules and minimize the impact of disruptions on travel plans. This predictive model can be used to improve operational efficiency, optimize resource allocation, and enhance customer satisfaction by providing more reliable delay estimates.
* Explain what data sources you are going to use.
* Specify your outcome variable, and the predictors you will use to predict the outcome.
* Describe your dataset: how many rows, how many columns, what types of variables are included? You can use the descriptive analytics lab as your basis and tweak it to suit your needs.
* Visualize some interesting relationships between a few predictors and the outcome to understand the relationships between them.

You should **avoid**:

* Time series prediction/forecasting: avoid datasets where the main predictor is the outcome variable at a prior time (e.g. stock prices where a key predictor of price today is price yesterday).
* Classification problems: make sure you outcome is numeric and continuous. The exception is problems where the outcome is a 0-1 indicator. You can treat these as regression problems.
* Small datasets: too few rows (say fewer than 100), or too few columns (say fewer than 10).
* Datasets with a lot of missing values.

This presentation is developmental and not evaluative; it does \*not\* count towards your final team grade.

**Datasets:**

| **Name** | **Title(s)/Link(s)** | **Description** | **Ideas** |
| --- | --- | --- | --- |
| **Uriel** | [**Flight Delay and Cancellation Dataset (2019-2023)**](https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023/data?select=flights_sample_3m.csv)  [Airline Delay and Cancellation Data, 2009 - 2018](https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data) | * Source: US Department of Transportation, Bureau of Transportation Statistics | **Problem**: predict whether a flight will be delayed based on factors like weather, airline, departure airport, time of day, etc.  **Outcome variable**: flight delay time (minutes)  **Predictors**: weather conditions, airline, departure/arrival airports, day of the week, flight distance, etc.  **Relevance**: airlines: optimization of schedules and improving customer experience. Customers: avoid flights that are predicted to be delayed |
| **Zac** | [**Microsoft Malware Prediction**](https://www.kaggle.com/competitions/microsoft-malware-prediction/data?select=train.csv)  [**Heart Disease Prediction**](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data) | * Criminy 8GB. Data from Microsoft * 14 attributes, 1000 observations from 1988 | **Problem:** Predict a Windows machine’s probability of getting infected by various families of malware, based on different properties of that machine.  **Outcome Variable:** Whether or not (1/0) machine detects malware  **Predictors:** 84 different characteristics about Windows computers including country, city, platform, processor, OS, bunch of other computer stuff I don't understand.  **Relevance:** Can be used to prevent computers from getting malware.  **Problem:** Predictthe presence of heart disease in the patient based on various medical factors  **Outcome Variable:** Has Heart disease (1/0)  **Predictors:** age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol, etc.  **Relevance:** Can be used to predict if/when a person has or is likely to have a heart disease and potentially take steps to reduce the likelihood of heart disease. |
| **Vivian** |  |  |  |
| **Albert** |  |  | **Flight data but only BOS Logan** |
| **Francisco** |  | **-Loan club database in Keggle** | **Problem:** Predictthe **loan default risk** by estimating the **potential financial loss** using borrower profiles and loan details. This is a continuous, numeric outcome rather than a binary classification.  **Outcome Variable:**  **potential financial loss**  **Predictors:**  Loan amount   * + Borrower’s annual income   + Credit score   + Employment status   + Loan purpose   + Debt-to-income ratio   **Relevance:** Can be used to predict risk of loan |