**ChBE 6745 – Group 6 Project definition**

**Github: https://github.com/yship1002/ChBE-DataScience-Group6**

**Hazardous materials incidents dataset – To develop a regression model to predict total damage cost with a** R² **score greater than 0.8**

**Background:**

The US pipeline and hazardous materials safety administration (PHMSA) collects detailed incident reports of hazardous materials (hazmat) transport accidents in the U.S. These reports contain structured data on incidents which includes the type of material, quantity released, package type, and damages incurred. From perspective of chemical engineering, understanding and mitigating financial loss due to hazmat incidents is important. By analyzing this data, we can better predict the financial severity of incidents based on incident characteristics – potentially improving packaging, transport safety, and risk mitigation strategies.

**Project goal:**

The goal of this project is to develop a regression model that predicts the financial severity of hazmat transportation incident, defined by total damages incurred. The model’s performance will be evaluated using R² score with 0.8 being our target.

### **Strategy**

Our strategy focuses on developing predictive models for estimating the total amount of damage resulting from hazardous material incidents. We will preprocess the dataset by aggregating multiple rows per incident, cleaning missing values. Categorical features such as transportation mode, hazardous class, and packaging type will be encoded, while continuous features like quantity released and estimated speed will be standardized. Initial modeling will use linear regression as a baseline, followed by non-linear models such as Random Forest, XGBoost, or maybe neural nets to capture nonlinear relationships and interactions between features. Model performance will be evaluated using standard criteria such as RMSE, MAE, and R².

### **Potential Risks**

Several risks may affect the project. First, data quality issues (missing or inconsistent entries, mixed data types) may limit model accuracy. Second, the distribution of damages is heavily skewed, with a small number of extreme-loss incidents that may dominate predictions; traditional regression may underestimate such rare but impactful cases. Third, high-cardinality categorical variables (e.g., carrier names, shippers) may lead to sparse encoding. Training and tuning tree-based models may require significant computation, particularly with high-cardinality categorical variables. Finally, advanced models like XGBoost or neural nets may offer higher accuracy but provide limited transparency for stakeholders who need explainability for safety or regulatory purposes.

**Dataset Description:**

For this project, we will use the Hazardous Materials Incident dataset obtained from the U.S. Department of Transportation’s PHMSA records. The dataset contains 10,237 incidents with 201 features describing event details, including transportation mode, location, hazardous material type, failure cause, and financial impacts. The target variable is Total Amount of Damages, which aggregates material loss, property damage, and cleanup costs. Preprocessing steps included handling missing values, aggregating multiple rows per incident, log-transforming the damage cost to address skewness, and applying one-hot encoding to categorical features to prepare the data for modeling.