Project Proposal for CS-GY 6053: Foundations of Data Science

Group Members: Akshat Singh (as20255), Abha Wadijkar (aw5399), Shweta Shekhar (ss19623)

1. Research Question

- Primary question: How do different weather conditions impact food delivery times across varying distances?
- This question is important for understanding how external factors impact delivery efficiency. By analyzing how weather conditions like storms or fog affect delivery times over various distances, companies can improve time estimates, optimize routes, and set clearer customer expectations.

2. Data Description

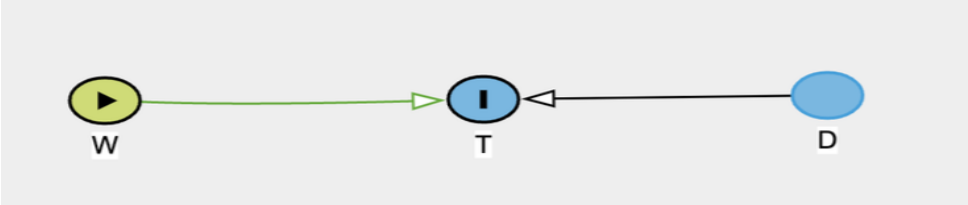
Source: [Kaggle - Food Delivery Dataset](https://www.kaggle.com/datasets/gauravmalik26/food-delivery-dataset)
Key Variables:

1. **Order ID:** Unique identifier for each food delivery order.
2. **Delivery Time (T):** Time taken for the delivery to be completed, measured in minutes. This is the primary outcome variable.
3. **Distance (D):** Distance between the restaurant and the delivery location.
4. **Weather Conditions (W):** Describes the weather at the time of the delivery, categorized as sunny, rainy, stormy, foggy, etc.
5. **Order Time:** The timestamp when the order was placed, which can be used to analyze the impact of time of day on delivery times.
6. **Delivery Location Type:** Indicates the type of location (e.g., urban, suburban, or rural), which can impact delivery routes and times.
7. **Traffic Conditions:** Traffic status at the time of delivery, if available, to study its influence on delivery time.

Sample Size:
The dataset consists of approximately **10,000** delivery records, ensuring enough data points for statistical analysis and model training.
Description:
The dataset captures various aspects of food delivery logistics, focusing on delivery times under different conditions. It provides detailed information on factors like weather and distance, enabling analysis to understand how external variables impact delivery efficiency. The data can be used to identify patterns and correlations between weather conditions and delivery time, informing potential operational adjustments for food delivery services.

3. Causal Model: Directed Acyclic Graph (DAG):

- Treatment Variable: W - Weather conditions (e.g., stormy, foggy, sunny)
- Outcome Variable: T - Time taken (min)
- Confound Variable: D - Distance between the restaurant and delivery location



4. Proposed Statistical Model

Model Type: Multivariate Linear Regression and Decision Tree Regression

Justification: We propose to use a combination of multivariate linear regression and decision tree regression to model the impact of weather conditions on delivery times across varying distances. The multivariate linear regression model is suitable for this analysis as it can quantify the linear relationship between the outcome variable (delivery time) and multiple independent variables (weather conditions and distance). This allows for understanding how each independent variable contributes to changes in delivery time. The decision tree regression model is chosen as a non-linear approach to capture complex interactions between weather conditions and distance that may not be well-captured by a linear model. Decision trees can handle non-linear relationships and provide interpretable outputs, making it easier to understand how different weather scenarios impact delivery times.
Potential Adjustments:

- **Interaction Terms:** We may introduce interaction terms between weather conditions and distance in the linear regression model to see how their combined effect impacts delivery times**.**
- **Feature Engineering:** Additional features such as time of day, traffic conditions, and road types may be considered if data is available or can be inferred.
- **Regularization Techniques:** If overfitting is detected, regularization techniques like Lasso or Ridge regression may be applied to improve model generalization.
- **Evaluation Metrics:** Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) will be used to evaluate the performance of the models.
- **Data Preprocessing:** Imputation of missing data, scaling of continuous variables, and encoding of categorical variables (e.g., weather types) will be performed to ensure model robustness.