Project Plan:
Deliver an approximately 2 page plan, in HTML or PDF format, addressing the following elements in order:

Project Title: “Chicago Collision Chronicles.” Alternatively, “Windy City Wreck Watch” or “Chi-Town Crash Chronicles.”

Group Members:
Murtaza Jawid, Kevin Norris, Yota Sugai, Nathan Young

Topic - Car crashes in Chicago

Questions of Interest: At least one for regression and one for classification. (We should add a couple backups)

- Regression Questions:
    - Can we build a regression model to predict the cost of damage based on factors such as posted speed limit, weather conditions, lighting conditions, and type of crash?
        - Backup Question 1: Can we predict the duration of road closures following a crash based on similar factors?
        - Backup Question 2: Is there a correlation between the severity of injuries sustained and crash characteristics?
        - Backup Question 3: Can we identify hotspots for crashes based on geographical location data (latitude and longitude) and other contributing factors?

- Classification Questions:
    - Can we classify whether a crash resulted in fatalities based on various attributes such as weather conditions, road defects, and traffic control devices?
        - Backup Question 1: Can we predict the severity of injuries (eg. minor, moderate, severe) in car crashes using similar attributes?
        - Backup Question 2: Is it possible to classify crashes based on the type of vehicles involved (e.g., passenger cars, trucks, motorcycles)?
        - Backup Question 3: Can we predict the variables most likely to predict Hit and Runs.

        - Backup Question 4: Can we predict the most likely locations for car accidents? Will our models have different predictions based on the severity of the crash?

Planned Approach: This includes framing the problem, the data, the methods, and the Intended Outcome.

- Framing Problem
- Framing Data
- Framing Methods
- Intended Outcome

Literature Review: Identify at least one article or publication per group member relevant to the topic. It does not have to be a scholarly article but it shall provide context for the topic and questions of interest to help in framing the analysis.

- https://www.sciencedirect.com/science/article/pii/S0001457518300873
- https://repository.rit.edu/cgi/viewcontent.cgi?article=12218&context=theses
- https://cyberlaw.stanford.edu/blog/2013/12/human-error-cause-vehicle-crashes
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7932803/

Initial Literature Review Findings:

- Factors that are believed to increase the risk of car accident include
    - Young age
    - Drugs and Alcohol
    - Cold temperature
    - Poor eyesight
- Factors that contribute to deadly and more dangerous car accidents
    - Speed limit
    - Icy conditions
    - Driving under the influence

Data Assessment:

The dataset consists of car crashes that occurred on city streets within Chicago between January 1, 2016 and December 31, 2023. It contains 48 variables and 818,218 rows, including information on damage, weather conditions, road defects and other crash-related attributes.

Planned Methods:

- Identification of the expected methods for both regression and classification.
- Identify planned approaches for assessing multicollinearity and model comparison as appropriate to the course.
- Identification of potential ethical concerns.
- Risk Assessment and mitigation.

- Regression Methods
    - Linear Regression: Predict damage cost based on numerical predictors such as speed limit and other conditions
    - Polynomial Regression: Account for potential non-linear relationships between predictors and damage cost
    - Decision trees: Capture interactions and non-linearities in the data to predict damage cost
- Classification Methods
    - K-Nearest Neighbors (KNN): Classify crashes as fatal or non-fatal based on similar instances in the dataset
    - Linear Discriminant Analysis (LDA): Separate fatal and non-fatal crashes based on linear combinations of predictors
    - Quadratic Discriminant Analysis (QDA) : Capture non-linear decision boundaries between fatal and non-fatal crashes

(These methods are based on our original questions – if we need to resort to additional questions or frameworks, we will modify the methodology outcomes accordingly.)

Deliverable:

We are going to produce a written report (approx. 8-10 pages in length) detailing our methods and findings. This report will discuss the data, our approaches to interpreting the data, and our recommendations for what to do to effectively answer our questions.

In addition, we will attempt collaboration on Github between the four of us in order to simulate a real-world collaborative project experience.

Schedule and Hours:

Week (Activities By This Date)	Activities	Estimated Hours	Expected Outcomes
April 1	• Discuss and finalize specific objectives for regression and classification analysis	10 hours per member	• Clearly defined regression and classification questions • Initial dataset exploration

	• Begin exploring the dataset to understand structure, potential, challenges and data cleaning requirements		with potential challenges identified
April 8	• Determine which variables/features are relevant for regression/classification tasks • Identify and discuss any ethical considerations related to the data and analysis	3 hours per member	• Finalized list of variables/features for regression/classification • Documented ethical considerations and mitigation strategies
April 15	• Data Cleaning and Modeling: Implement chosen methods, assess multicollinearity, compare models • Risk Assessment: Identify potential risks in the analysis process and develop mitigation plans	8 hours per member	• Cleaned dataset ready for analysis • Implemented regression/classification models with initial evaluation • Documented risks and mitigation strategies
April 22	• Conduct through evaluation of regression/classification models • Create appropriate plots and diagrams to illustrate model results and insights • Begin preparies the poster presentation/project report	8 hours per member	• Comprehensive evaluation of model performance • Visualizations demonstrating key findings • Progress made toward final presentation/report
April 25	• Finalize and polish the report for clarity and coherence • Ensure each member has completed their assigned tasks and contributions • Prepare all required materials for submission according to the guidelines of the project	CRUNCH TIME (think Bethesda or Naughty Dog levels of crunch)	• Completed and polished project report • Individual contributions aligned with group responsibilities • Submitted project ready for assessment

Group Member Responsibilities

Responsibilities to be divided:

- Data Cleansing
- Variable and model selection
- Regression Analysis
- Classification Analysis
- Report Outline
    - Drafting
- Red team
    - Assess multicollinearity
    - Variable overlap
    - Identification of potential ethical concerns.
    - Risk Assessment and mitigation

Group Member	Responsibilities
Murtaza Jawid	Red team, Regression Modeling
Kevin Norris	Red team, Report Outline, Classification Modeling
Yota Sugai	Data Cleaning, Variable and model selection, Classification Modeling
Nathan Young	Data Cleaning, Report Outline, Regression Modeling

Summary

This statement confirms that the project plan is meticulously formatted, presenting information clearly and concisely while minimizing errors. Clarity, precision and minimal mistakes are imperative for effective communication and successful project execution.