Revised Project Plan: Accident Severity Analysis and Prediction

Objective:

Analyze accident data to gain insights into patterns across geographical locations using **EDA** and **visualizations**. Develop a **model** that predicts accident severity, with outputs displayed on **Folium maps** for deeper geographical insights.

Action Plan: Step-by-Step Breakdown

1. Dataset Loading

- Import necessary libraries.
- Load the dataset (ensure it covers **only data from 2019 and beyond**).
- Quick **data overview** to understand columns and structure (.info(), .head()).

---

2. Exploratory Data Analysis (EDA)

- **Visualize missing values** with heatmaps.
- **Distribution of severity levels** using bar charts.
- **Analyze accident trends**:
  - **Accidents by time** (hour, day of the week, month).
  - **Weather and accident severity** relationships.
- **Folium-based geographical analysis**:
  - Accident hotspots by state and city.
  - Use scatter maps to plot the distribution of accidents.

Plots to Include:

- Bar charts for severity distribution.
- Time-based histograms (Hour, Day, Month).
- Box plots for temperature and visibility against severity.
- **Folium scatter maps** for accident locations.

---

3. Data Cleaning

- **Filter only 2019+ data** (since earlier data is out of scope).

- Handle **missing values**:
  - Drop columns with too many missing values (e.g., End_Lat, End_Lng).
  - Impute missing weather data with **median/most frequent value**.
- **Remove duplicates** and **negative durations** (if any).
- Ensure all **categorical data is encoded** appropriately for model training.

---

4. Feature Engineering and Preprocessing

- Create new **features** for better model performance:
  - **Rush hour**: 7-9 AM, 4-6 PM.
  - **Weekend** indicator.
  - **Bad weather indicator** (e.g., rain, snow, fog).
- Encode categorical variables using **one-hot encoding** or **label encoding**.
- Scale continuous variables using **StandardScaler** if needed.
- **Ensure at least 24 features** are present in the final dataset.

---

5. Model Training and Validation

- **Train-test split** (80/20) on the cleaned dataset.
- **Model selection**:
  - Train an **XGBoost classifier** for severity prediction.
- **Evaluate model**:
  - Report **accuracy, precision, recall**, and **F1-score**.
  - Visualize confusion matrix and classification report.

---

6. Prediction Output with Folium Maps

- Use the trained model to **predict accident severity** for test data.
- Plot the **predicted severity levels** on a **Folium map** with:
  - **Color-coded markers** (Red for high severity, Green for low severity).
  - Include **hover-over information** (city, time, weather).

7. Presentation and Final Output

- Display **all relevant visualizations** in the notebook.
- Provide **conclusions and insights** from the analysis.
- **Discuss** the limitations of the model and potential improvements.

Project Deliverables

1. **Jupyter Notebook** containing:
    - Dataset loading, EDA, cleaning, preprocessing, and model training.
    - Folium maps showcasing predicted accident severity.
    - Detailed explanations and visualizations.
2. **Graphs and Maps** for EDA:
    - Severity trends and distribution.
    - Weather and time patterns.
    - Accident prediction results plotted on a **Folium map**.