**Part 1: Categorical/Linguistic Analysis**

**1. Stakeholder Category/Text Analysis Needs:**

**Stakeholder Categories & Key Insights**

1. Law Enforcement & Traffic Safety

* Focus: Crash causes, time-based trends, high-risk locations.
* Needs: Compare trends, filter by severity, map accident hotspots.
* Text Analysis: Keyword frequency, topic grouping, sentiment on injury severity.

2. City Planners & Transportation Officials

* Focus: Infrastructure issues, urban crash hotspots, seasonal patterns.
* Needs: Visualize accident-prone areas, analyze road design issues.
* Text Analysis: Topic modeling for urban planning issues, geospatial clustering.

3. Public Health & Emergency Services

* Focus: Injury severity trends, emergency response needs, behavioral patterns.
* Needs: Track injury rates, compare severity by cause, assess crash risks.
* Text Analysis: Sentiment scoring, narrative categorization.

4. Insurance & Risk Assessment

* Focus: Crash risk factors, driver behavior analysis, severity patterns.
* Needs: Compare high-risk behaviors, predict future trends.
* Text Analysis: Keyword analysis for liability, topic modeling on crash causes.

**Text Data Type**

* Structured Categories: Collision Type, Injury Type, Primary Factor.
* Semi-Structured Text: “Primary Factor” allows NLP-based analysis.
* Analysis Focus: Broad themes (human, environmental, mechanical) & specific topics (e.g., speeding, failure to yield).

2. **Data Assessment:**

Data Structure

* Categorical Data: Predefined categories like Collision Type, Injury Type, Primary Factor.
* Textual Data: "Primary Factor" includes structured categories but some have additional explanations.
* Mixed Data: Categorical, numerical (Year, Hour, Latitude/Longitude), and free-form text (Reported Location).

Quality & Consistency Issues

* Missing Values: "Primary Factor" (1,121), "Hour" (225), "Weekend?" (68), among others.
* Inconsistencies: "Primary Factor" contains explanations; needs standardization.
* Geospatial Data: Some missing Latitude/Longitude may affect mapping.

Preprocessing Plan

1. Handle Missing Data: Fill/drop missing values based on context.
2. Standardize "Primary Factor": Remove extra explanations & group similar causes.
3. Process Locations: Aggregate common crash sites.

**3. Initial Design Exploration:**

I'll create two different visualizations for categorical and textual insights.

1. Bar Chart: Frequency of Primary Crash Factors

Why? Helps stakeholders understand the most common causes of accidents.

How? Group "Primary Factor" categories and visualize frequency.

2️. Word Cloud: Crash Causes

Why? Provides a quick, high-level view of dominant accident-related keywords.

How? Use NLP techniques to extract and visualize key terms.

Here are two different visualizations to explore categorical and textual patterns in the dataset:

1. Bar Chart: Top 15 Causes of Car Accidents

- This visualization highlights the most frequent primary factors contributing to crashes.

- Helps stakeholders (e.g., law enforcement, policymakers) identify key risk factors and prioritize safety measures.

A graph of a graph

AI-generated content may be incorrect.

2. Word Cloud: Crash Causes

- Provides a quick high-level view of dominant accident-related keywords.

- Useful for identifying recurring terms in crash causes.

- Can be refined further by removing stop words and grouping similar factors

A close-up of words

AI-generated content may be incorrect.

**Part 2: AI-Assisted Design Process**

**1. Documenting AI Interactions**

* AI Model Used: ChatGPT (OpenAI, GPT-4)
* Prompts Used:
  + - "Analyze this dataset for categorical and linguistic insights."
    - "Suggest meaningful visualizations for accident data."
    - "Generate Python code to visualize categorical and textual data, including bar charts and word clouds."
    - "Summarize the data assessment concisely."
* **Why These Prompts?**
* Concise & Direct: Focused on specific tasks (analysis, visualization, summarization).
* Step-by-Step Refinement: Broke down tasks into data assessment, preprocessing, visualization, and best practices.

**2. Implementation Plan**

* **Data Preparation Steps:**
  + Handle Missing Data: Fill/drop missing values in "Primary Factor", "Weekend?", "Hour", and geospatial fields.
  + Standardize Textual Data: Clean "Primary Factor" to remove explanations and merge similar categories.
  + Process Location Data: Aggregate common accident sites for better pattern analysis.
* **Data Analysis & Visualization Tools:**

A screenshot of a computer

AI-generated content may be incorrect.

* **Interactive Features:**
  + Filter by Time: Allow users to analyze crashes by year, month, and hour.
  + Geospatial Heatmaps: Interactive maps for high-incident locations.
  + Crash Factor Comparison: Allow stakeholders to compare different crash causes across injury types.
* **Handling Data Quality Issues:**
  + **Drop or Impute Missing Data**: If missing values are significant, use median/mode imputation or remove incomplete rows.
  + **Standardize Categories**: Merge similar accident causes for consistency.
  + **Remove Unnecessary Text in "Primary Factor"**: Focus on key causes, removing explanations.

**3. Evaluating AI Suggestions**

* **Helpful Contributions:**
  + Drop or Impute Missing Data: If missing values are significant, use median/mode imputation or remove incomplete rows.
  + Standardize Categories: Merge similar accident causes for consistency.
  + Remove Unnecessary Text in "Primary Factor": Focus on key causes, removing explanations.
* **Limitations Encountered:**
  + - Drop or Impute Missing Data: If missing values are significant, use median/mode imputation or remove incomplete rows.
    - Standardize Categories: Merge similar accident causes for consistency.
    - Remove Unnecessary Text in "Primary Factor": Focus on key causes, removing explanations.
* **Modifications and Improvements:**
  + Drop or Impute Missing Data: If missing values are significant, use median/mode imputation or remove incomplete rows.
  + Standardize Categories: Merge similar accident causes for consistency.
  + Remove Unnecessary Text in "Primary Factor": Focus on key causes, removing explanations.