Project Brief: Distracted Driving Impact Analysis: Fault & Severity

Step 1: Determine Your Topic and Data Source

- **Topic:** I am investigating the impact of **Distracted Driving** on traffic accidents in Montgomery County, MD. Specifically, I want to determine if distracted drivers are more liable for accidents and if they cause more severe damage.
- **Question:** Does distracted driving significantly increase the likelihood of a driver being determined "At Fault," and does it lead to more severe vehicle damage compared to non-distracted drivers?
- **Alignment:** This aligns with the "common good" requirement by addressing public safety and driver behavior.

Step 2: Acquire and Document Your Data

- **Source:** I acquired the **"Crash Reporting - Drivers Data"** from Data.gov, which is maintained by Montgomery County, MD.
- **Unit of Analysis:** Each row in my dataset represents a single **driver** involved in a traffic collision.
- **Key Variables**:
  - Driver Distracted By: My independent variable (Distracted vs. Not Distracted).
  - Driver At Fault: My first dependent variable (Binary: Yes/No).
  - Vehicle Damage Extent: My second dependent variable (Ordinal scale 0-4).
- **Cleaning:** I filtered out "Unknown" distraction statuses, standardized text columns to uppercase, created a binary is_at_fault column, and mapped damage descriptions (e.g., "DISABLING") to numeric scores.

Step 3: Formulate Your Hypothesis

I used the "Only One Test" framework to set up my hypothesis:

1. **Test Statistics:**
   - **Difference in Proportions:** For the fault rate.
   - **Difference in Medians:** For the vehicle damage severity score.
2. **Null Hypothesis:**There is no difference in the proportion of "At Fault" drivers or the median vehicle damage severity between distracted and non-distracted drivers.
3. **Alternative Hypothesis::** Distracted drivers are significantly more likely to be found "At Fault" and will have a higher median vehicle damage score.

Step 4: Identify Metrics for Bootstrap Uncertainty

I selected two metrics for bootstrapping:

1. **Metric 1 (CLT Applies):** Difference in Proportions (Fault Rate).
    - *Why:* Proportions generally follow a normal distribution with large samples.
2. **Metric 2 (Non-CLT Metric):** Difference in **Median** Vehicle Damage Score.
    - *Why CLT does not apply:* The damage score is discrete and ordinal (0-4 scale), and the sampling distribution of a median is generally not normal, making standard parametric tests inappropriate [1111].

Step 5: Conduct Your Analysis

In my notebook, I carried out these steps:

- **Loaded and Cleaned:** I handled missing values and standardized the distraction and fault columns.
- **Permutation Test:** I ran **5,000 permutations** for the Fault Rate metric, resulting in a p-value of **0.0**, which strongly rejects the null hypothesis.
- **Bootstrapping:** I generated **5,000 bootstrap samples** to create 95% confidence intervals for both metrics.
    - *Fault Rate CI:* [0.656, 0.663] (Distracted drivers are ~66% more likely to be at fault).
    - *Median Damage CI:* [0.0, 1.0] (Inconclusive lower bound, but potential for higher severity).

Step 6: Interpret and Communicate

My final analysis tells the following story:

- **Question:** I asked if distraction leads to fault and damage.
- **Pattern Found:** I found a **massive statistical gap** regarding fault—distraction is a definitive predictor of causing an accident. However, the *severity* of the damage (median) is less conclusive, with the confidence interval including 0.
- **Certainty:** I am extremely certain about the fault attribution (p-value 0.0) but have identified nuance/uncertainty regarding the damage severity.
- **Conclusion:** Distracted driving guarantees a higher risk of *causing* an accident, but it does not guarantee the accident will be more *severe* than a non-distracted crash.