Updated Topic Check IN

For my project, I am investigating the impact of driver distraction on both fault determination and crash severity. Specifically, I hypothesize that drivers reported as distracted (e.g., texting, eating) are significantly more likely to be deemed "at fault" and will generally be involved in crashes with higher median vehicle damage compared to non-distracted drivers. I will be analyzing the **"Crash Reporting - Drivers Data"** provided by Montgomery County, Maryland. Using this dataset of over 150,000 clean records, I will perform a permutation test to compare fault rates and use bootstrapping to estimate the uncertainty of the difference in **median vehicle damage scores**, satisfying the project requirement to analyze a metric where the Central Limit Theorem (CLT) does not apply.

Here are the specific changes I made to the topic check-in:

1. I expanded the project scope to include **crash severity** (measured by median vehicle damage) alongside fault determination, which incorporates the required second metric.
2. I explicitly specified that I will use bootstrapping on **median vehicle damage scores**, satisfying the requirement to analyze a metric where the Central Limit Theorem (CLT) does not apply.
3. I updated the dataset size from "over 200,000" to **"over 150,000 clean records"** to accurately reflect the volume of data remaining after my cleaning process.

Project Scaffold Table (Week 12 Check-In)

Element	**Your Plan**
Topic/Question	**Distracted Driving Impact Analysis:** 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?
Hypothesis	**Alternative Hypothesis (H_1):** Distracted drivers are significantly more likely to be found "At Fault" and will have a higher median vehicle damage score than non-distracted drivers.
Outcome/Metric/Test Statistic	1. **Difference in Proportions** (Fault Rate): The difference in the percentage of drivers found "At Fault" between groups. 2. **Difference in Medians** (Damage Severity): The difference in the median "Vehicle Damage Score" (0-4 scale) between groups.

Element	Your Plan
Units of Analysis	Each observation represents a **single driver** involved in a reported traffic collision in Montgomery County, MD.
Data Source(s)	**"Crash Reporting - Drivers Data"** from Data.gov (Montgomery County, MD). URL: https://catalog.data.gov/dataset/crash-reporting-drivers-data
Why this data works	The dataset is large (~160,000+ rows) and contains explicit columns for Driver Distracted By (the independent variable), Driver At Fault (outcome 1), and Vehicle Damage Extent (outcome 2), making it perfect for direct group comparison.
Uncertainty Metric	**Difference in Median Vehicle Damage Score.** I will use bootstrapping to estimate the 95% confidence interval for this metric because the Central Limit Theorem does not apply to the sample median, and the sampling distribution of the sample median is generally not normal (even in large samples) for discrete or ordinal distributions.
Null Hypothesis	**Null Hypothesis (H_0):** There is no difference in the proportion of "At Fault" drivers or the median vehicle damage severity between drivers who are distracted and those who are not.