Project 2 Proposal: Analyzing NYC Collision Data

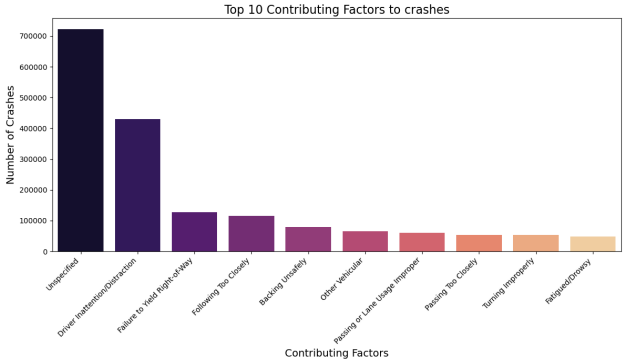
Team Members: Yu-Sheng Lee, Jeremy Cui, Hannah MacDonald
Github Repository:
https://github.com/UC-Berkeley-I-School/project2_lee_cui_macd

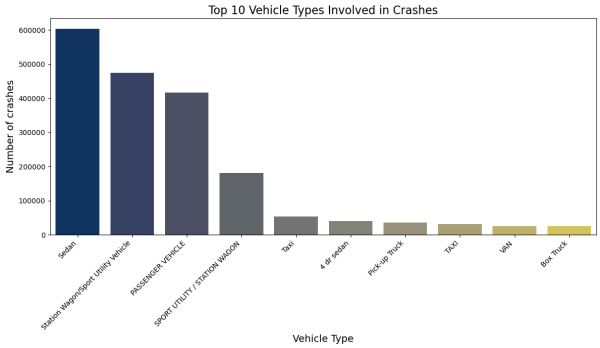
Datasets:

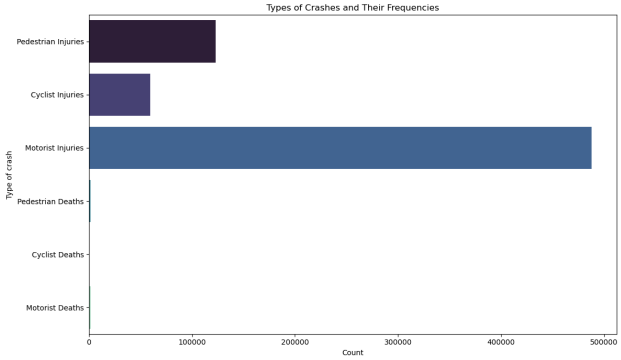
- Primary: NYC OpenData Motor Vehicle Collision - Crashes dataset describes one crash event that is recorded by the NYC police department in each row. All collisions are recorded when someone is injured or killed, or when there is at least $1000 worth of damage.
- Secondary: Property Valuation and Assessment Data

Initial Plots, Figures, or Tables

The top contributing factors to crashes are: driver inattention/distraction. The vehicle types most likely to be involved in a crash are sedans and station wagons/SUVs. Motorist injury has the highest frequency of a crash. We will investigate collision data further by referencing geographical and socioeconomic factors and provide insights in the final deliverable.







Variables of Interest

Our team aims to conduct exploratory data analysis to explain the factors of the crashes.

- ON STREET NAME (Primary dataset):
    - What: Name of street on which incident occurred
    - Why: In order to cross-reference the location of the crash with the property value of the community it occurred in
    - Field on secondary dataset: AV-STREET-NAME

- VEHICLE TYPE CODE #:
    - What: Type of vehicle involved in the incident
        - Values include: ATV, bicycle, car/suv, ebike, e-scooter, truck/bus, motorcycle, other
    - Why: In order to answer if specific vehicle types lead to more incidents

- AV-FINAL-AVT-ACT:
    - What: Final actual assessed total value of the property
    - Why: In order to infer the socioeconomic condition of the unit of area

Main Research Question: Do underrepresented communities in New York City experience a greater magnitude of traffic incidents?

1. What vehicle types, time in a day, and locations are most likely to be involved in a crash and why?
2. How does the severity of the crash distribute?
3. How is underrepresented communities defined?
    a. Should we define it by gross income, % of minorities, house prices, etc.?
4. What vehicle types lead to the most casualties?

Final Deliverable

A comprehensive report detailing our findings, supported by statistical analysis and data visualizations. An in-class presentation will summarize the key insights and methodologies.