16-791 Applied Data Science
Project Proposal

Team name: Squirrel Hill Research
Project Title: Park Safe
Team members:

- Aashi Gupta (aashig)
- Gautham Kumar Vedam (gvedam)
- Vivek Gupta (vivekgup)

1. GOAL

The goal of this project is to study and characterize the vehicular thefts in San Francisco city during the period 2018-2020. We aim to perform this study along the dimensions - time and location demographics of the theft incident. We further plan to use the insights derived from this analysis in conjunction with SF Parking Meters data to recommend safe parking spots to drivers provided the time & their location.

2. MOTIVATION

San Francisco city is the cultural, commercial, and financial center of Northern California. The city is plagued by a car break-in epidemic with 25,677 car break-in reports in 2019, according to data from the San Francisco Police Department (source). The police department only makes arrests in about 1% of cases. Vehicle thefts shot up almost 21% in San Francisco from April to May in 2020, compared with the same period last year, in spite of nearly non existent traffic due to Coronavirus.(source)

Identifying safe parking spots will assist the city locals and tourists looking to park their vehicles. In a broader context, it may also help people looking to purchase or rent property in a safer neighborhood.

3. DATA

- **Parking Meters** data published by the Municipal Transportation agency of SF, which includes a list of all parking meters in SF (source)
- **Police Department Incident Reports: 2018 to Present** data published by SF Police Department, is the main dataset for vehicle thefts (source)
- **Income by Location in San Francisco** data provided by US Census (source)
- **Population by Neighborhood in San Francisco** data by US Census (source)

4. HYPOTHESES

- Vehicle thefts are more likely to occur during wee hours of the day
- Poor neighborhoods have a higher rate of vehicle thefts
- More populous areas experience lower crime rates

5. TECHNICAL APPROACH

Using the `"Incident Category"` column of **Incident data**, we filter out vehicle theft records. Next, we can use `"Incident Datetime"` & `"point"` (location) to perform initial exploratory analysis. Appending the **Income data** and **Population data** can further enable us to test other hypotheses. We can then build a predictive model which given the time and location outputs the probability of vehicle theft. Finally, we use **Parking Meter data** to suggest safe parking spots.

6. EXECUTION RISKS & MITIGATION STRATEGIES

We identify a risk in constructing the (binary scoring) predictive model using the **Incident data.** Since the dataset is a Crime Report, all the data points belong to the same category (*positive = crime*). Using this imbalanced data will result in an impractical model which always outputs the same class. To mitigate this risk, we intend to artificially add *negative* data points, i.e. no crime incidents, for every dataset location reported periodically at an empirically decided interval. We will test for 10 *minute*, 1 *hour*, *etc.*

We also need to map location coordinates to neighborhoods to suggest safe parking spots. We will use appropriate location to area mapping libraries (like *geopy*) to mitigate this risk.