General Assembly
Data Science – Potential Final Project Ideas
Tony Leung
January 14, 2017

Idea #1

Problem statement

San Francisco continues to experience population growth as the economy grows. As there is an influx of wealth into the City, crime increases. One area of crime is vehicle break-ins. Using Jan 2016 through Dec 2016 data (eg. break-in category, Day of Week, Date, Time, Incident Address, location, population density, wealth of each neighborhood) published from the City of San Francisco, determine the likelihood of a vehicle break-in at any particular location in San Francisco.

Hypothesis

A car is broken into: early mornings, during the week, a relatively higher income neighborhood.

Potential dataset:

https://data.sfgov.org

Idea #2

Problem statement

There are many restaurants in San Francisco and they are subject to food inspections to protect the public. Using food inspection scores on kaggle.com, determine if a restaurant will eventually be closed from bad scores (alternatively predict when a restaurant will be closed).

Hypothesis

A restaurant will shut down from bad food inspections if it receives multiple bad inspection scores, the severity of such violations is high, and the restaurant does not resolve critical violations within the allotted time.

Potential dataset:

https://www.kaggle.com/datasf/sf-restaurant-inspection-scores

Idea 3

Problem statement

As flying becomes more popular in the United States, airline delays are expected, as coordination among airlines and passengers become more complex. Using the US Dept of Transportation data collected of large air carriers (via kaggle.com), determine the amount of time an airline is delayed.

Hypothesis

An airline will be delayed if: weather was bad, departure time is later in the day, flying out from a hub, and at busy times of the week (Monday, Thursday, Friday).

Potential dataset:

https://www.kaggle.com/giovamata/airlinedelaycauses