MATH 448 Project Proposal: Understanding Public Data and Predict Probability of Success of Newly Opening Business.

- Team :
  Jongwook Choe (no other team members).

- Description of the Problem :
  Haven't you worried about stepping on a poop while walking down the road? Well, I did. "San Francisco's public poo problem hits record level. They call it the Bay Area brownout. San Francisco authorities have recorded more than 125,000 cases of human feces found on public streets since 2020." (Washington Times) According to the article, just the cases reported in 2020 are more than those reported from 2011 to 2019. I believe it started after the covid hit. The city is starting to get better but we want clean shoes. I thought if we opened a shoe cleaning business in a neighborhood where there is a higher chance people shoes it could be a big hit!

- Description of the Data:
  In this project, we will use two main data resources

  - 311 Cases DataSF
    San Francisco keeps all manner of data about its street, including where the city gets the most calls about poop.

    https://www.sf.gov/

    https://data.sfgov.org/City-Infrastructure/November-2014-Street-Cleaning-TTC/p39p-ttts/about_data

    This dataset includes cases generally associated with a place or thing (for example parks, streets, or buildings) and created July 1, 2008 or later. Cases generally logged by a user regarding their own needs (for example, property or business tax questions, parking permit requests) are not included

  - Yelp API
    https://docs.developer.yelp.com/reference/v3_business_search
    returns up to 1000 businesses with some basic information based on the provided search criteria in this case it will be shoe cleaning.

- Methodology:
  Throughout the analysis, the goal is to predict where will be the optimizing place for the new opening business. The data will be trained by merging public data and Yelp API. I will create a new equation to estimate the success rate based on review counts and counts of surrounding human waste.

- Possible Challenge:
  How to weigh columns for the success rate can be very subjective therefore I will start with EDA to see the relationship between variables such as the coefficients and correlation.