

# Customer Risk Classifier

## Introduction

Uses US traffic accident dataset gathered over a 3-year period (2016 ~ 2019) and a dataset of customers' information (e.g. their street address and average miles driven per day) to categorize a customer as a high-risk or low-risk client, providing grounds for increasing or decreasing his/her insurance premium.

## Assumptions

1. For most customers, they do not travel outside of their state for their daily commute
2. A circular area determined by the range of a customer's daily commute provides a reasonable stretch of land that the customer will dwell in for a large majority of his or her time
3. The number of accidents that occur within a given area over a certain period of time is indicative of how often accidents occur within that area.
4. There is a direct correlation between a customer's likelihood of getting into an accident and the regularity of collisions that occur within his or her mile radius of daily travel

## Hypothesis

Therefore, a customer's risk level can be measured by the regularity of accidents that occur within a given mile radius of a customer's commute and the location in which he or she resides

## Procedure

1. Use Google Maps API to convert each customer's home address to its longitudinal and latitudinal coordinates
2. For each accident that occurs within a customer's state, calculate the distance of that accident from the customer's residence by using the customer's previously calculated geographical coordinates and the coordinates of the accident (provided by the US traffic dataset)
3. Count the number of accidents whose distance from the customer's residence fall within the radius of his/her daily commute
4. After gathering the number of accidents per customer (hereafter described as Accidents Index) sort the data according to the Accidents Index in ascending order and use the median to split the data in two.
5. Repeat the last step 2 more times, further splitting the data into 8 zones (see below) (median was used instead of average since average is sensitive to outliers and especially influential to a small dataset such as a dataset of 30 customers being categorized)
6. For each customer, determine which zone the customer's Accidents Index falls, and use this to categorize each customer's risk level:

|                    |                     |
|--------------------|---------------------|
| 1 = Very Low       | 5 = Average         |
| 2 = Low            | 6 = Moderately High |
| 3 = Moderately Low | 7 = High            |
| 4 = Average        | 8 = Very High       |
7. Use Google Sheets API to output the customer classification data to a desired spreadsheet document

