# Customer Segmentation for Insurance Dataset

# Basic Idea of Project

- Read through an insurance dataset that has various variables that describe each customer, some of which are missing / empty for different customers.
- Clean and analyze the data to make recommendations about several clustered types of customers

# Data Cleaning Process

- The existence of NaNs meant that various techniques to deal with them were explored including:
- Dropping those rows / columns (depending on the frequency of NaNs for a specific variable);
- Using a nearest neighbor technique to find similar customers (from the other variables) and copy the missing value from that similar customer;
- Using averages to fill in missing values.
- After correcting the NaN data, the existence of outliers required those rows to be either dropped or the whole column / variable normalized.

# Preprocessing

- After fixing NaNs and outliers, the data was preprocessed. This included running correlation analysis to determine variables that were too closely correlated to be independent of each other. Too much data width creates sparsity and makes machine learning models more difficult to train effectively.

# Analysis

- The remaining variables were then analyzed to see if distinct clusters of customers could be created.
- The techniques used include: Quartiles, Box Plots, Decision Trees, Groupby tables, K-Means, Meanshift, Gaussian Mix, DBScan, and K-Modes

# Skills Learned

- Working with Insurance data with realistically missing data.
- Cleaning and preprocessing data
- Applying and tuning various machine learning models