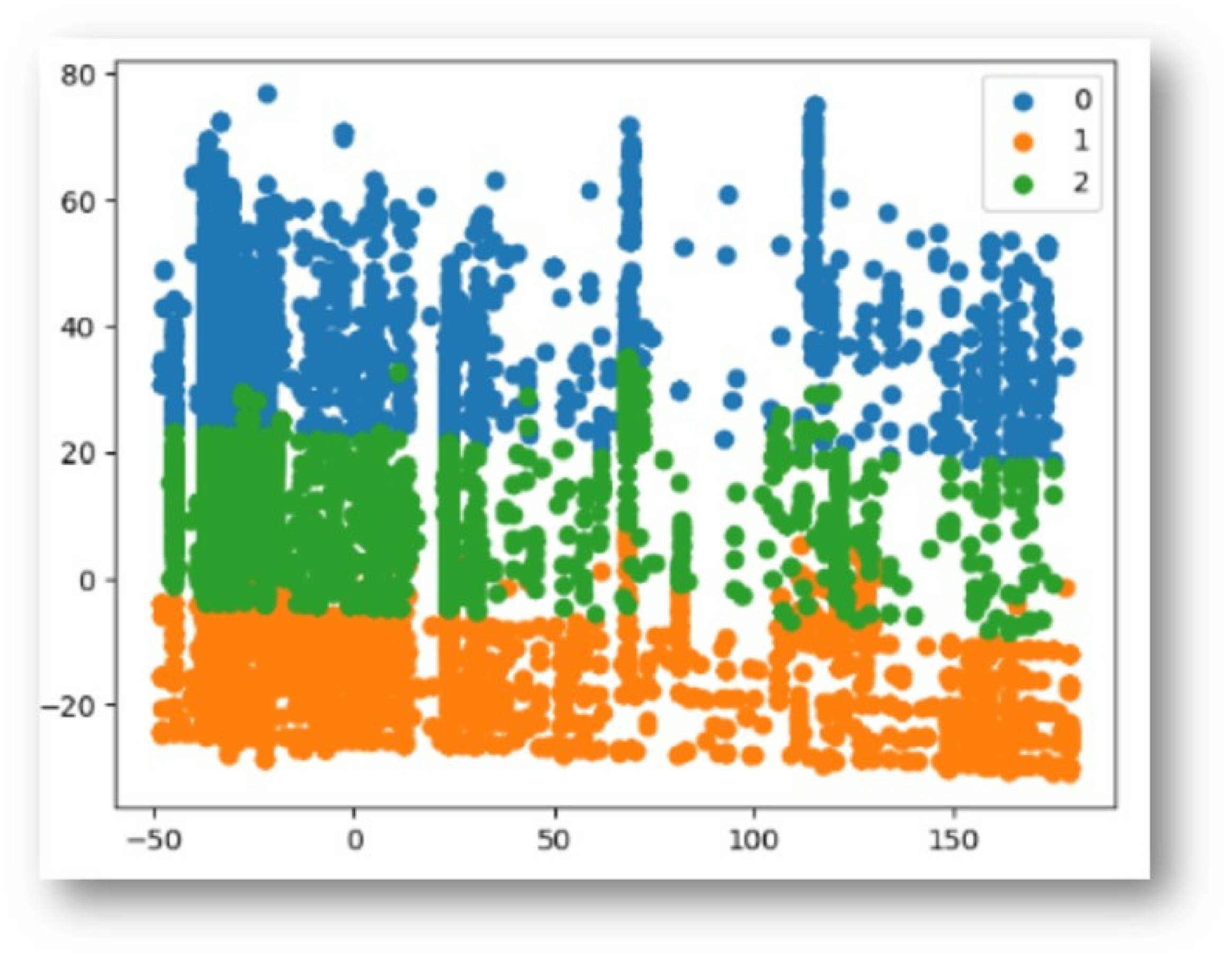
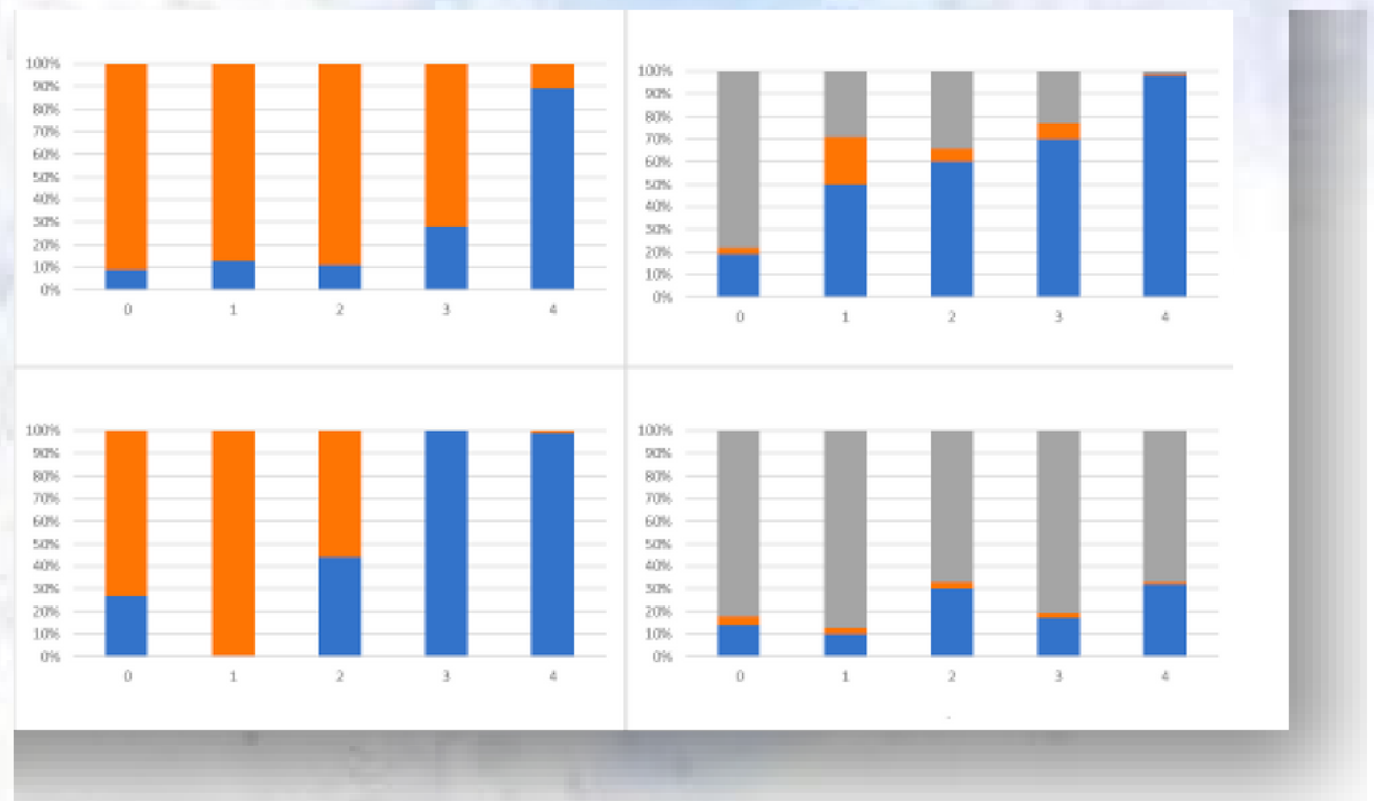
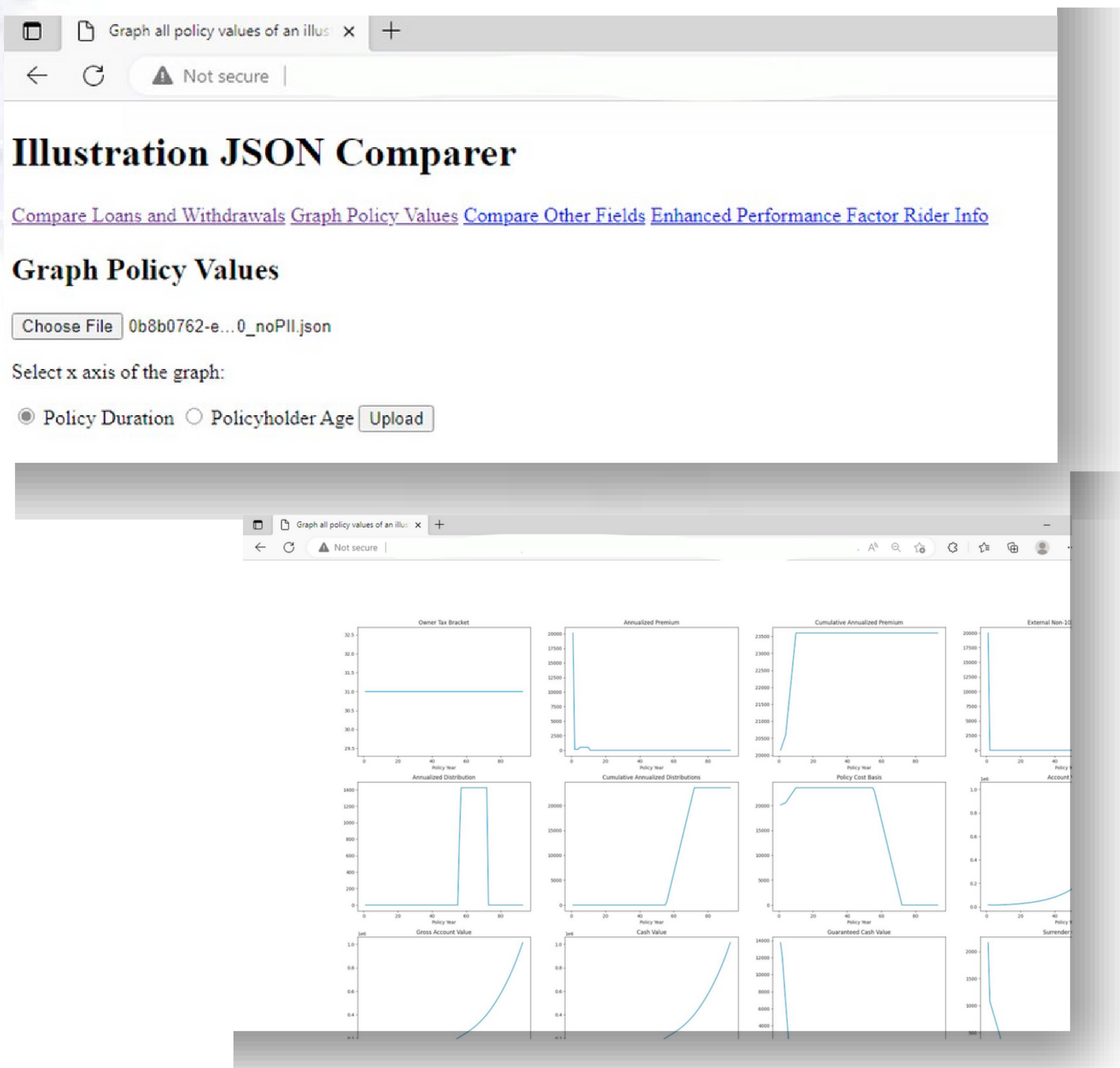
CUSTOMER SEGMENTATION WITH ML & AI



The distribution across the centroids (or the clusters). Each color is a different cluster. The y-axis indicates the deviation from the median value. We can see that there are some significant outliers.



Data from clusters can be represented in stacked bar charts like the ones shown above. Each cluster has slightly different proportions of features enabled.



Graph all policy values of an illus

Not secure

Illustration JSON Comparer

Compare Loans and Withdrawals Graph Policy Values Compare Other Fields Enhanced Performance Factor Rider Info

Graph Policy Values

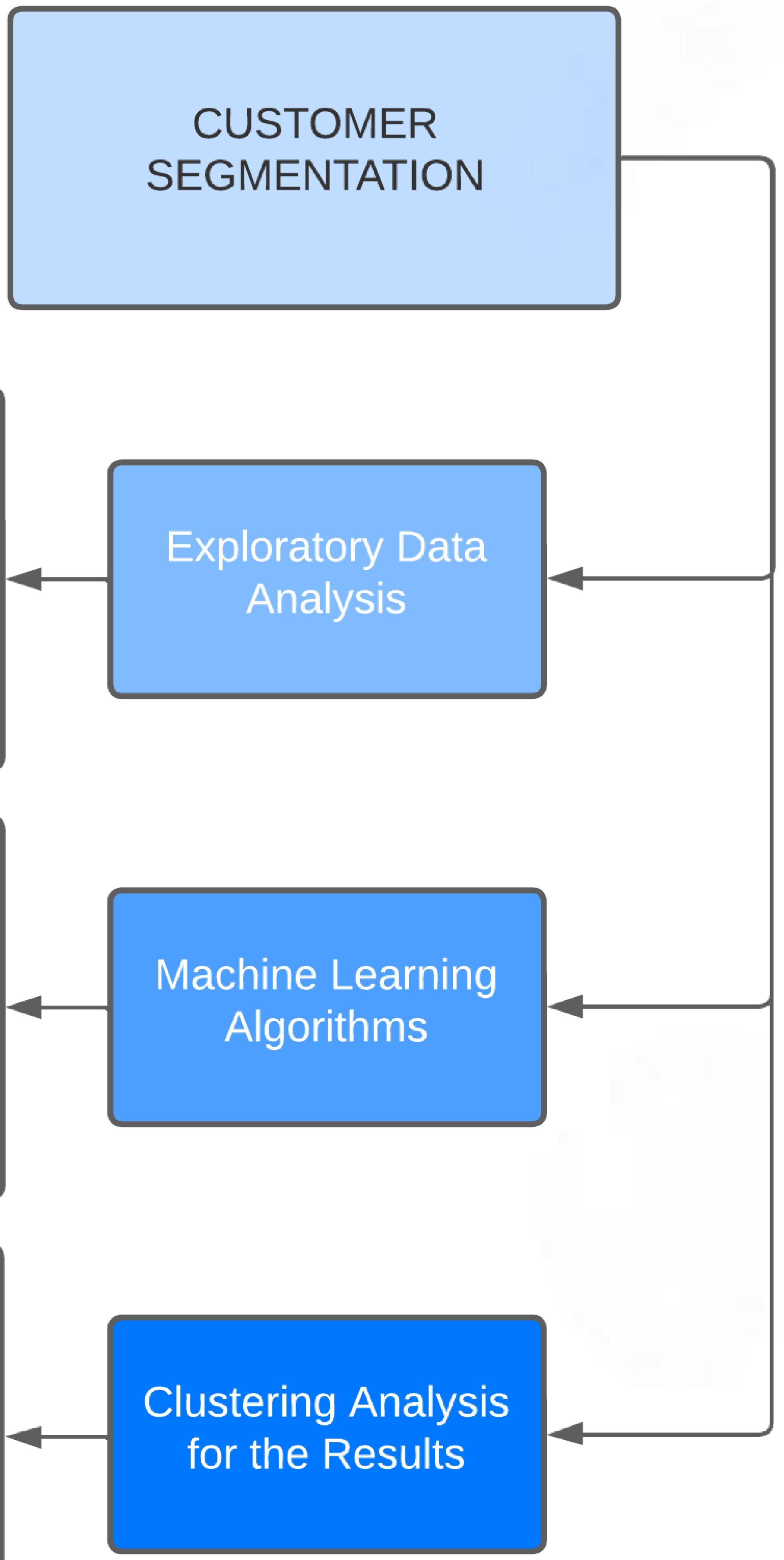
Choose File 0b8b0762-e...0_noPII.json

Select x axis of the graph:

● Policy Duration ○ Policyholder Age Upload

Graph all policy values of an illus
Not secure
Owner Tax Bracket
Annualized Premium
Cumulative Annualized Premium

The json-comparing webapp created to visualize data files.



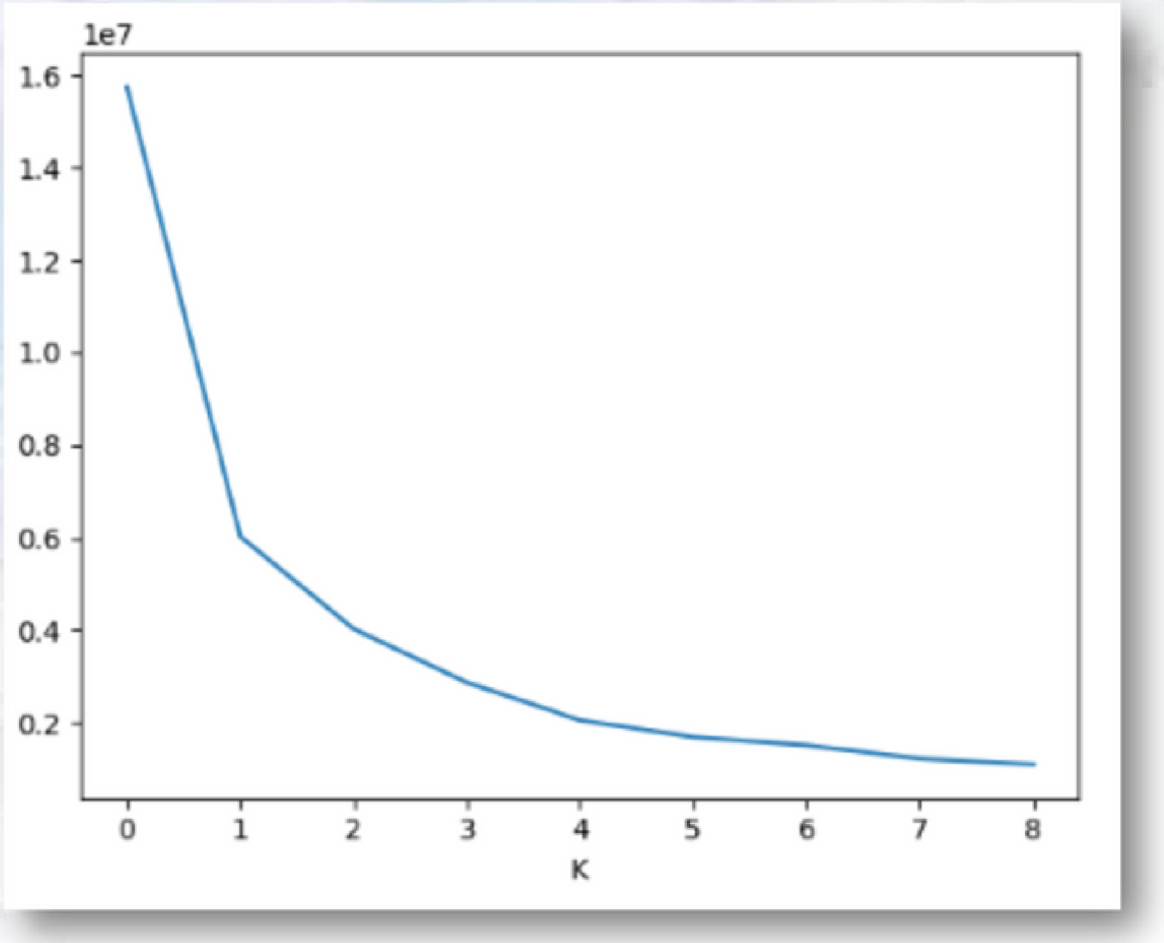
CUSTOMER SEGMENTATION

Exploratory Data Analysis

Machine Learning Algorithms

Clustering Analysis for the Results

Exploratory Data Analysis, or EDA for short, is the process of performing initial investigations on a given dataset. Our teams would first attempt to spot anomalies within the dataset, which is typically indicated by a drastic jump in the policy face amount or an irregular policy plan pairing. Then divide the data into specific groups so that the outputs produce meaningful results. Then, we would try to produce reasons for the discrepancies, construct stories for why people may choose to make their policies this way, and report our findings.



The x-axis describes the number of clusters, varying from 0 to 8, and the y-axis is the "score" for each k, which is computed as the sum of square distances from each point to its assigned center.

In order to maximize the effectiveness of our research, we need to spot connections that are too difficult to see manually. That’s why we apply various machine learning algorithms. As we are given tens of thousands of JSON files, our top priority was to process these data into something more manageable: Using pandas, we converted all these files into a single CSV. Applying the KNN algorithm, we could estimate values to replace missing information. Finally, utilizing a combination of K-prototype and the Knee-locator programs, we could form groups of policies with varying characteristics. Each of the “clusters” shares unique similarities, which can be further examined to find powerful connections.

We can choose which variables we want to focus on during the clustering by artificially raising the “importance” value of that said variable. If we wanted to measure the impact of LTC (a certain type of feature) on the average premium amount, we could raise the LTC’s value and observe the premium average change, thus determining whether these two variables are closely connected. Additionally, by creating smaller groups prior to the cluster formation, we can get a more detailed output for the data. For example, if we restrict the data stream to only a certain product, we can then apply the clustering techniques and find connections within the product only. We also created a locally-hosted web application with Flask to graph the differences between 2 JSON files. Given enough time for development, the goal for this program is to compare any 2 data streams at the user’s convenience.

PURDUE UNIVERSITY®

PACIFIC LIFE