Vehicle Insurance Policy

Group 3 - Jon, Giffin, Canxiu, Charles, Stephen

1. **Business Objective**
2. **Analytical Objective**
3. **Data Understanding**
4. **Data Exploration**
5. **Setup/Data preparation**
6. **Modeling**
7. **Evaluation**
8. **Deployment**

Business Objective

Background: The insurance company realizes that it does not have categories to label its customers and wants to create customer profiles (High risk customers, High-Middle customers, Low-Middle and Low customers). The thinking is that once customer profiles are created, the marketing team can then find new customers based on our team's results and market to them appropriately.

Goal: To segment customers who have the same annual_premium into different risk groups to design a more dynamic pricing approach where risky drivers pay more, and less risky drivers pay less.

Task: Build a clustering model that profiles and groups existing customers based on their common characteristics

Analytical Objective

Business problem: Currently, the company charges the same for each policy premium, but the company wants to move towards a more dynamic pricing approach where risky drivers pay more, and less risky drivers pay less.

Analytical solution: Apply unsupervised machine learning to cluster customers who have the same annual_premium into different risk groups by using personal, vehicle and financial data.

Machine learning method: We will build K Means model in PyCaret with 12 input features in order to cluster customers into 4 groups

...more to come

Data Understanding

Context:

The data contains the customers who have insurance policies with the same amount of annual premium affected during January to September 2020

Dataset (60393 instances):

Features:

- 4 Personal features, 5 Vehicle features and 3 financial features for each policy
- Feature engineering : Calculate Customers' age by using Policy Effective Date and Date of birth
- The following features are excluded for modeling: Policy Number, Annual Premium and claim office

Data Exploration

9 Categorical features: pol_number, pol_eff_dt, gender, date_of_birth, agecat, area, veh_age, veh_body, claim office

6 Numerical features: credit_score, traffic_index, veh_value, numclaims, claimcst0, annual_premium

Data quality assessment:

- pol_number: unique ids
- annual_premium: constant value
- date_of_birth: high cardinality (very high number of distinct values)
- pol_eff_dt: high cardinality
- agecat: 4873 (8%) missing values
- credit_score: 2801 (4.6%) missing values
- traffic_index: 3503 (5.8%) missing values
- claim_office: 50362 (83.4%) missing values
- numclaims: 50362 (83.4%) zeros
- claimcst0: 50362 (83.4%) zeros

Setup/Data preparation

1. Training and Testing split 90/10
2. Hyperparameters
    a. Ethical Issues
        i. Gender and Age
    b. Normalize
    c. Numeric Imputation
    d. Ordinal and High Cardinality Features
    e. Group feature

Modeling

Hierarchical Modeling Using Birch Clustering:

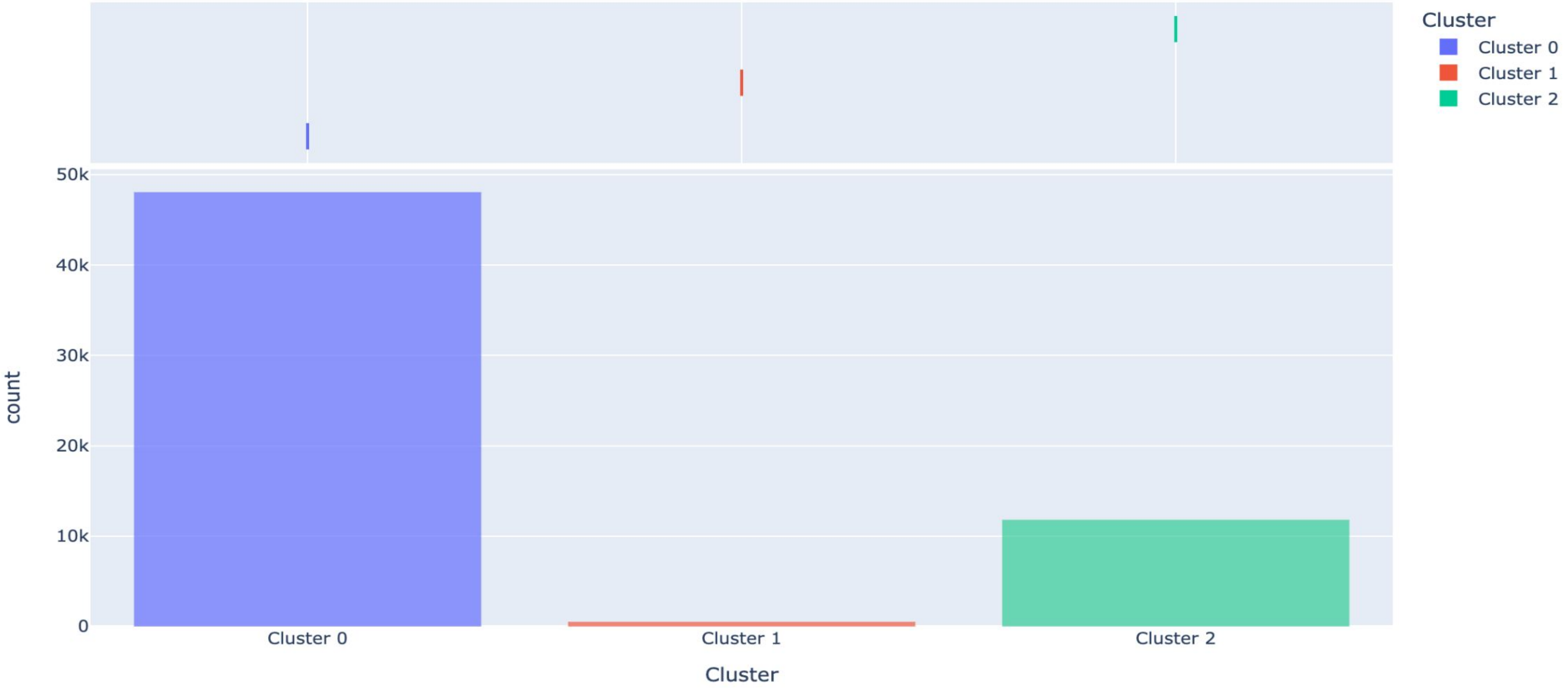
- Our birch model parameters: K=3 clusters, threshold=.70 and branching_factor=75

- Max recursion depth was reached using default threshold=0.5 and branching_factor=50

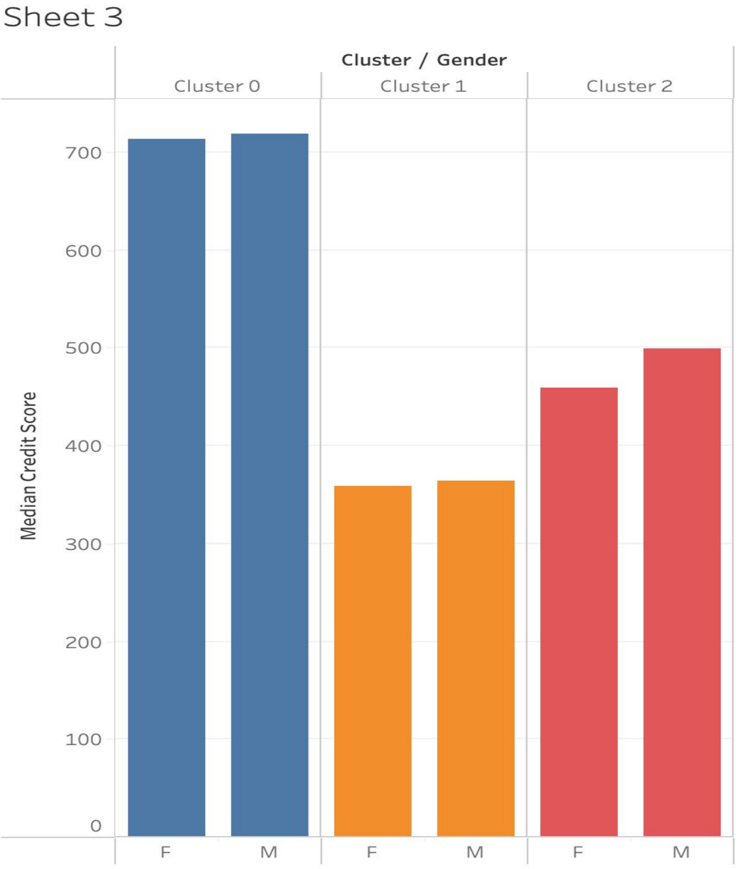
- Silhouette score was 0.143 and was much lower than our Kmeans model

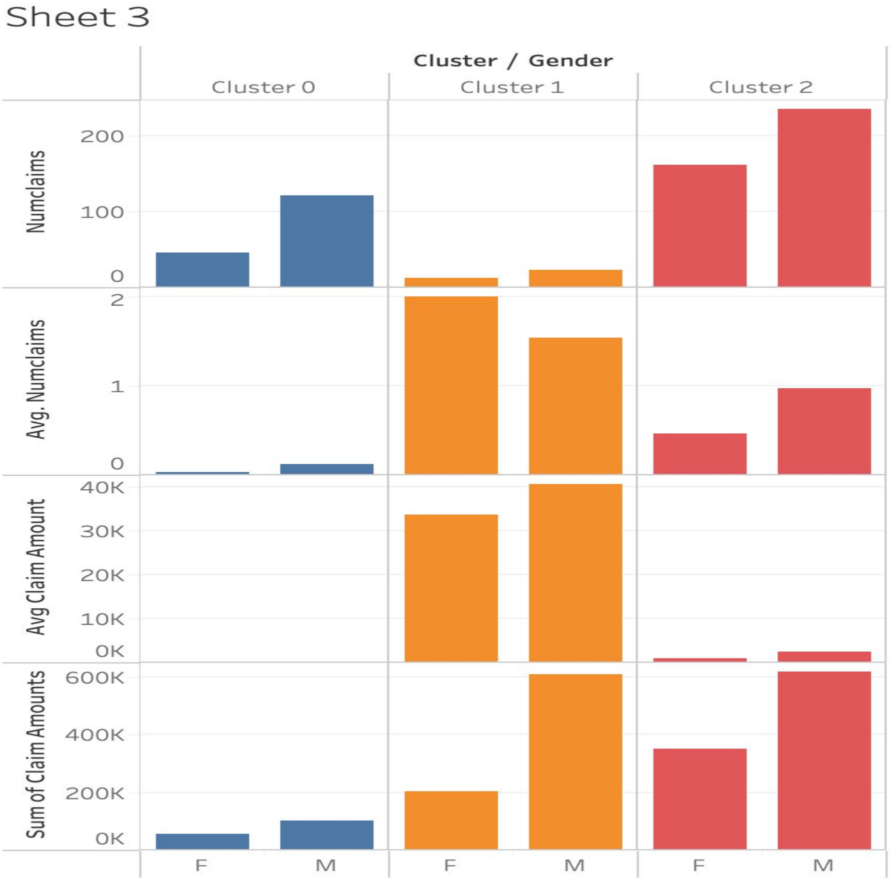
- Cluster distribution: Cluster 0 = 48k rows; Cluster 1 = 515 rows; Cluster 2 = 11.8k rows

Cluster distribution Plot for Birch Clustering:

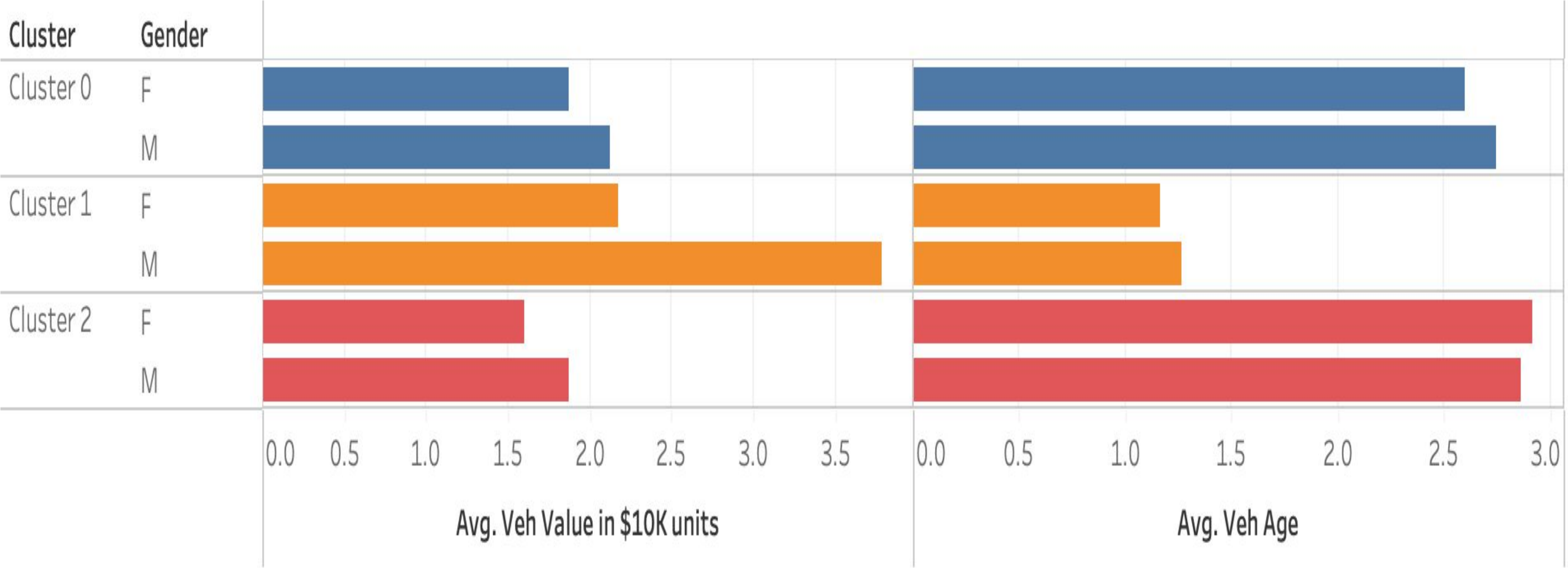


Birch Clustering Insights using Tableau Visualizations on Predicted Data for K=3 Clusters:





Birch Clustering Insights using Tableau Visualizations on Predicted Data for K=3 Clusters:



Clustering Insights For the Birch Hierarchical Model:

- Using Tableau to interpret the clustering results: Cluster 0 showed low-risk clients; Cluster 1 showed the highest risk clients; and Cluster 3 showed medium-risk clients.
- Cluster 1 clients had the lowest average credit score; Cluster 0 had the highest avg credit score
- Cluster 1 showed the highest avg number of claims and the highest average claim amounts
- Cluster 0 showed the lowest avg number of claims and the lowest average claim amounts
- Cluster 1 had the highest avg vehicle values and the lowest average vehicle age

Modeling

Partitioning Modeling: K-prototype Clustering

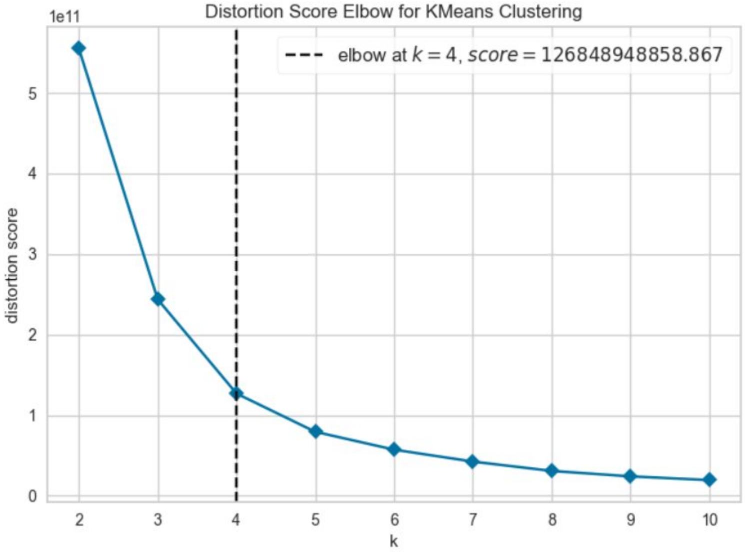
- K-prototype is able to handle categorical and numerical mixed features
- Calculate dissimilarity based on Euclidean distance and Gower distance
- Silhouette score was 0.108883
- Cluster distribution:
    - cluster 0: 20905
    - cluster 1: 11813
    - cluster 2: 21537

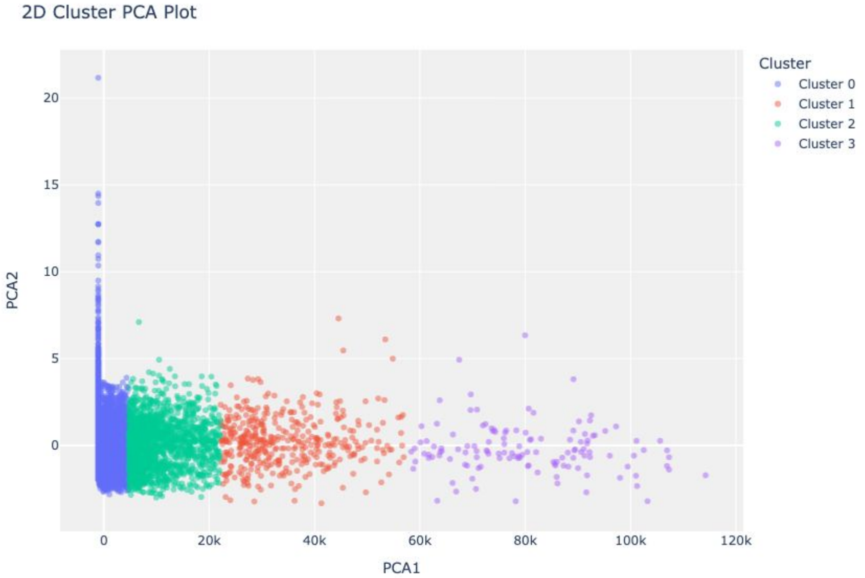
Winning Model

Partitioning Modeling: K-Means Clustering

- Combined claim cost and claim number as group features
- Removed unnecessary categorical features
- Silhouette score was 0.9338
- Cluster distribution:
    - cluster 0: 57714
    - cluster 1: 457
    - cluster 2: 2076
    - cluster 3: 145

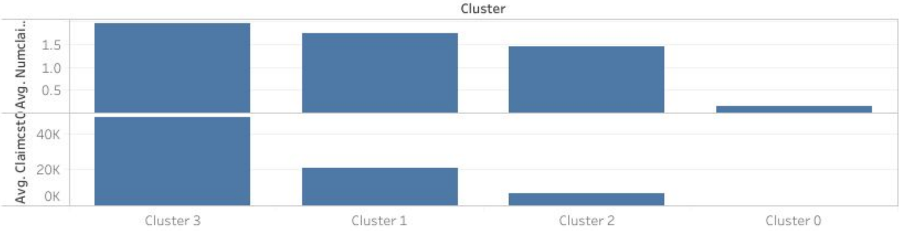
Evaluation


Distortion Score Elbow for KMeans Clustering
1e11
elbow at $k = 4$, $score = 126848948858.867$
distortion score
0
1
2
3
4
5
k
2
3
4
5
6
7
8
9
10

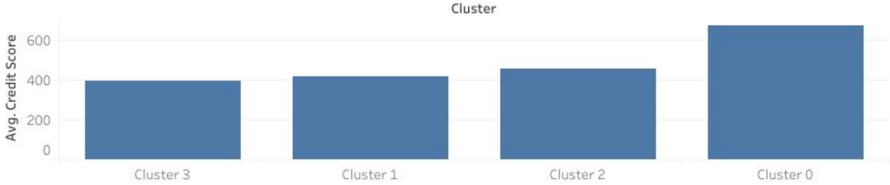

2D Cluster PCA Plot
Cluster
Cluster 0
Cluster 1
Cluster 2
Cluster 3
PCA2
0
5
10
15
20
PCA1
0
20k
40k
60k
80k
100k
120k

K-Means Clustering Analysis

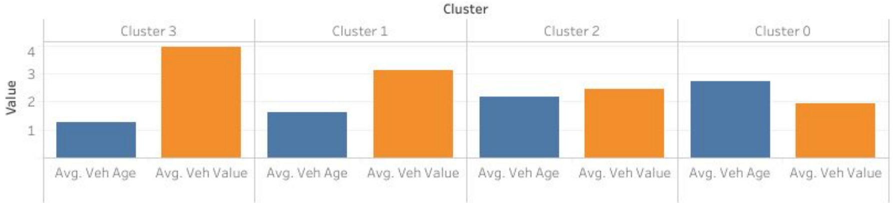
cluster vs. risk level



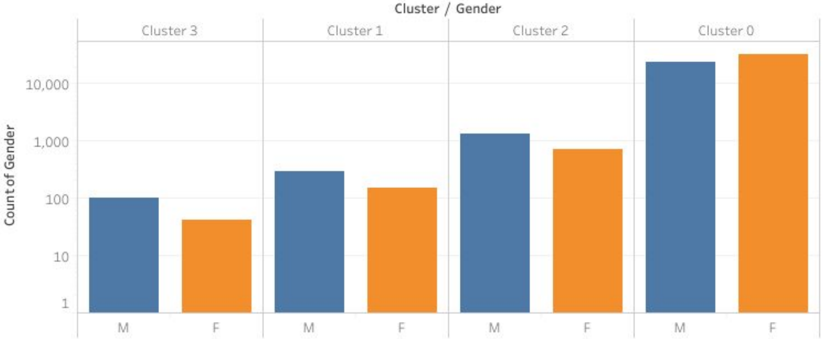
Does credit score affect the risk level?



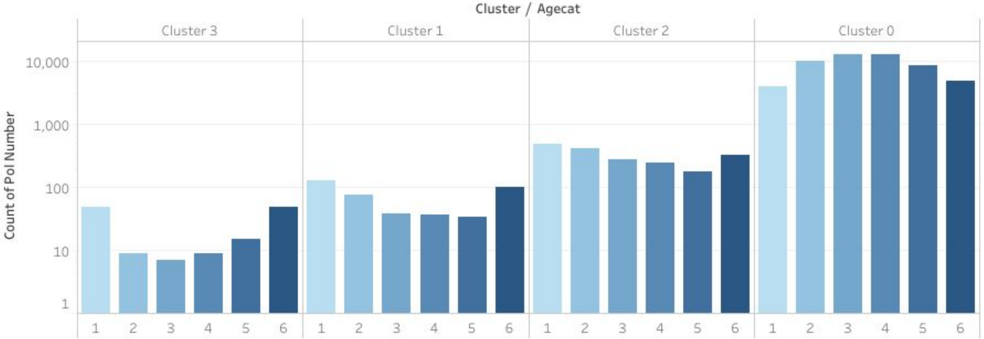
Does Vehicle affect the risk level?



Does gender affect risk level?



Does age cat affect risk level?



Deployment

- Styling
  - Bootstrap was used to make the app scalable as well as to ensure it was user friendly across devices/browsers
- Validation
  - Date Pickers
  - Number Entry
  - Text Entry
- Hosting
  - Uploaded the files to github
  - Deployed via Heroku

https://insurance-premiums.herokuapp.com