**Written Report: The Ability to Forecast Health of Customers**

Company XYZ is a health insurance company. They provide services and packages for employers to offer to their employees. To properly plan for each patient’s premium, Company XYZ has decided to build a program that will allow them to sort their new customers information into clusters. This information contains various demographic, health history, and lifestyle information (see data dictionary below with the factors included).

**Data Source**

This ‘new customer’ data was attained from Kaggle from the CDC's BRFSS2015. This data is real-world as it comes from 2015 CDC Behavioral Risk Factor Surveillance System public survey data. See link below for data source:

https://www.kaggle.com/datasets/prosperchuks/health-dataset?resource=download

**Data Dictionary**

|  |  |
| --- | --- |
| Age | 13-level age category (1 = 18-24, 2 = 25-29, 3 = 30-34, 4 = 35-39, … 13 = 80+ |
| Sex | 0 = Female, 1 = Male |
| HighChol | High cholesterol, 0 = No, 1 = Yes |
| CholCheck | Cholesterol checked in past 5 years, 0 = No, 1 = Yes |
| BMI | Body Mass Index |
| Smoker | 0 = No, 1 = Yes |
| HeartDiseaseorAttack | 0 = No, 1 = Yes |
| PhysActivity | 0 = No, 1 = Yes |
| Fruits | 0 = No, 1 = Yes |
| Veggies | 0 = No, 1 = Yes |
| HvyAlcoholConsumption | 0 = No, 1 = Yes |
| GenHlth | 1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair, 5 = Poor |
| MentHlth | Days of poor mental health in past 30 days, scale 1 - 30 |
| PhysHlth | Physical illness/injury days in past 30 days, scale 1 - 30 |
| DiffWalk | Difficulty walking, 0 = No, 1 = Yes |
| HighBP | High blood pressure, 0 = No, 1 = Yes |
| Stroke | 0 = No, 1 = Yes |
| Diabetes | 0 = No, 1 = Yes |

**The Approach**

In order to perform the clustering on this dataset, the data was normalized and the Kmeans() algorithm was used to sort the data into appropriate clusters. Using Kmeans() with the Euclidean distance is the most appropriate option with sensitive healthcare information, because the company wishes to be fair with its customers. Correlation is not causation and they do not wish to penalize if there may be correlations between spurious factors.

**The Results**

This cluster analysis produced three clusters of customers with varying trends. These three clusters seem to capture three different segments found that customers fit into. These three different clusters are named as follows based on the general trend of the factors analyzed:

Cluster 1: Middle-Aged & In Poorer Health

Cluster 2: Younger & In Better Health

Cluster 3: Older & In Moderate Health

Chart, line chart

Description automatically generated

Figure : Normalized Trends for Each Cluster

A more in-depth analysis of each cluster is provided below:

**Cluster 1:**

|  |  |
| --- | --- |
| Age | 0.244 (Middle) |
| Sex | -0.138 (More female) |
| HighChol | 0.360 (High cholesterol) |
| CholCheck | 0.048 (Tend to have checked cholesterol in past 5 years) |
| BMI | 0.473 (Highest BMI) |
| Smoker | 0.144 (Most Smokers) |
| HeartDiseaseorAttack | 0.667 (Most heart disease & attack history) |
| PhysActivity | -0.590 (Least physical activity) |
| Fruits | -0.078 (Least fruits) |
| Veggies | -0.094 (Less veggies) |
| HvyAlcoholConsumption | -0.224 (Least heavy alcohol consumption) |
| GenHlth | 1.108 (Worst general health) |
| MentHlth | 0.768 (Most poor mental health days) |
| PhysHlth | 1.411 (Most physical illness/injury days) |
| DiffWalk | 1.317 (Most difficulty walking) |
| HighBP | 0.502 (Higher blood pressure) |
| Stroke | 0.614 (Most history of strokes) |
| Diabetes | 0.608 (Most diabetes) |

**Cluster 2:**

|  |  |
| --- | --- |
| Age | -0.470 (Youngest) |
| Sex | -0.292 (More female) |
| HighChol | -0.551 (Lowest cholesterol) |
| CholCheck | -0.139 (Tend not to have checked cholesterol in past 5 years) |
| BMI | -0.386 (Lowest BMI) |
| Smoker | -0.160 (Least Smokers) |
| HeartDiseaseorAttack | -0.384 (Least heart disease & attack history) |
| PhysActivity | 0.182 (More physical activity) |
| Fruits | -0.058 (Less fruits) |
| Veggies | -0.098 (Least veggies) |
| HvyAlcoholConsumption | 0.083 (Most heavy alcohol consumption) |
| GenHlth | -0.647 (Best general health) |
| MentHlth | -0.158 (Less poor mental health days) |
| PhysHlth | -0.451 (Least physical illness/injury days) |
| DiffWalk | -0.458 (Least difficulty walking) |
| HighBP | -0.796 (Lowest blood pressure) |
| Stroke | -0.215 (Lowest history of strokes) |
| Diabetes | -0.785 (Least diabetes) |

**Cluster 3:**

|  |  |
| --- | --- |
| Age | 0.345 (Oldest) |
| Sex | 0.406 (More male) |
| HighChol | 0.362 (Highest cholesterol) |
| CholCheck | 0.119 (Tend to have checked cholesterol in past 5 years) |
| BMI | 0.110 (Higher BMI) |
| Smoker | 0.079 (Some Smokers) |
| HeartDiseaseorAttack | -0.018 (Less heart disease & attack history) |
| PhysActivity | 0.186 (Most physical activity) |
| Fruits | 0.068 (Most fruits) |
| Veggies | -0.045 (Less veggies) |
| HvyAlcoholConsumption | 0.057 (More heavy alcohol consumption) |
| GenHlth | -0.021 (Moderate general health) |
| MentHlth | -0.328 (Least poor mental health days) |
| PhysHlth | -0.426 (Less physical illness/injury days) |
| DiffWalk | -0.361 (Less difficulty walking) |
| HighBP | -0.534 (Highest blood pressure) |
| Stroke | -0.167 (Less history of strokes) |
| Diabetes | 0.454 (More diabetes) |

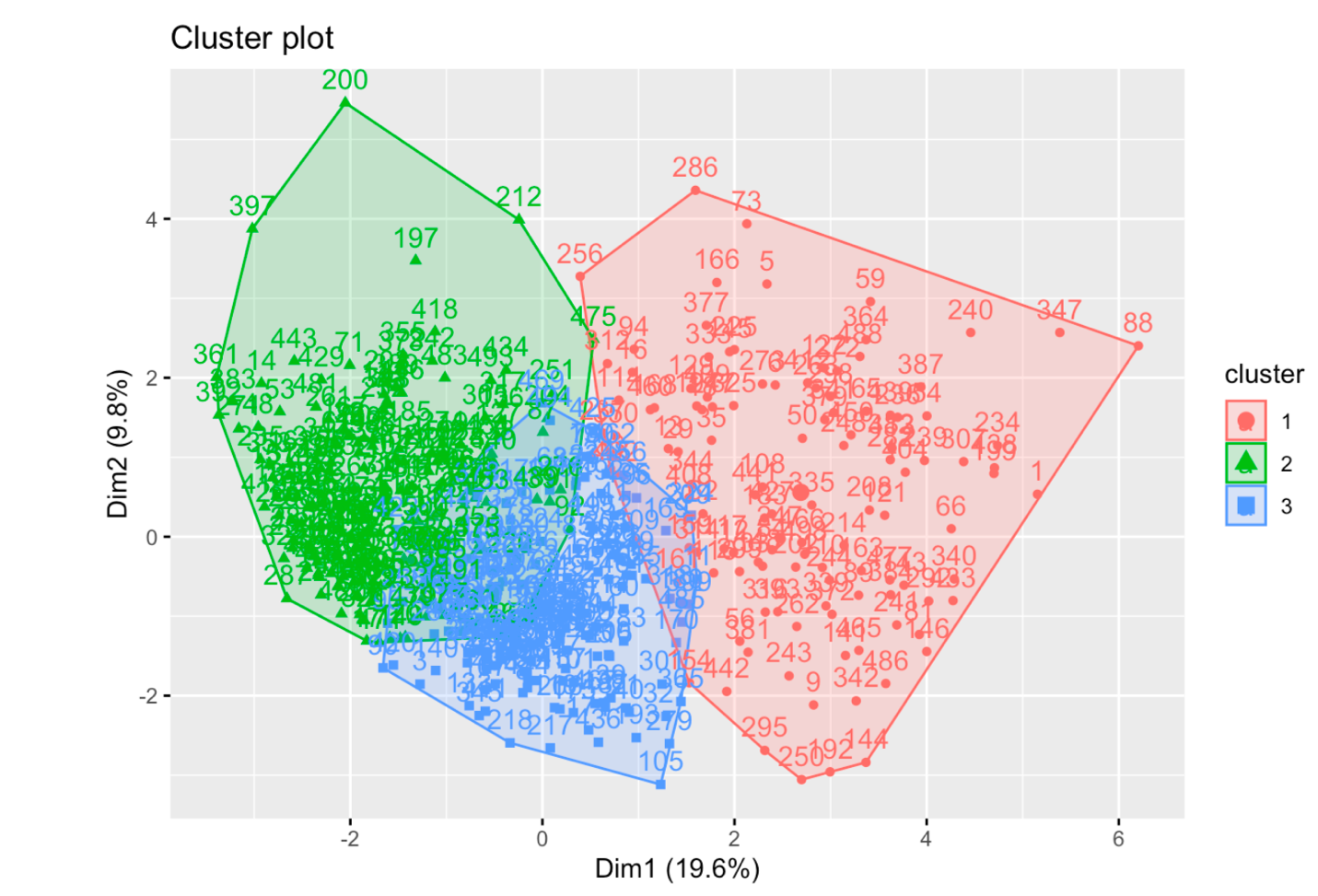


Figure Cluster plot for the 3 clusters identified

**The Conclusion**

Based upon the results of this analysis of their customers, Company XYZ is better able to classify and cluster in their new customers to appropriately forecast their health insurance premium needs. This system allows for the company to better cater to its customers and anticipate their costs accordingly. This ensures that the insurance company have sufficient premiums accrued when it needs to pay for healthcare services used by their customers. This clustering system also allows for the burden of payment to be spread more evenly across customer type.