RISK CLASS SEGMENTATION ON INSURANCE DATA

By Renee Chebet - 095919

PROBLEM STATEMENT

Analyzing Insurance claims data in order to segment customers into different risk classes based on their characteristics for purposes of premium calculation and creation of claim reserves.

DATA SOURCE

- The data was obtained from the database of an insurance company (Anonymized).
- It contains 581 rows and 9 columns.

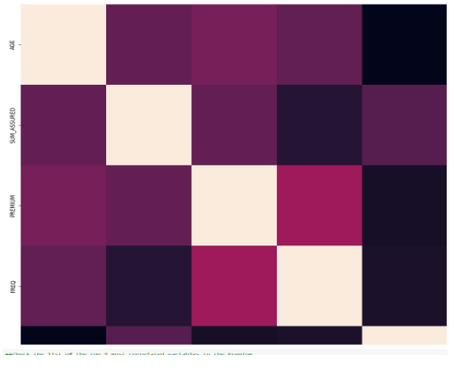
VARIABLES

Numeric Variables	Categorical
1. Plan Number 2. Age 3. Sum Assured 4. Premium 5. Frequency 6. Term	1. Gender 2. Claim Type

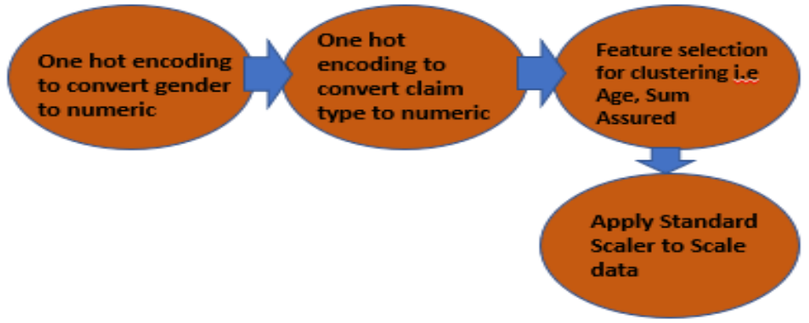
FEATURE SELECTION

The two most correlated features as highlighted by the correlation matrix are:

1. Age
2. Sum Assured

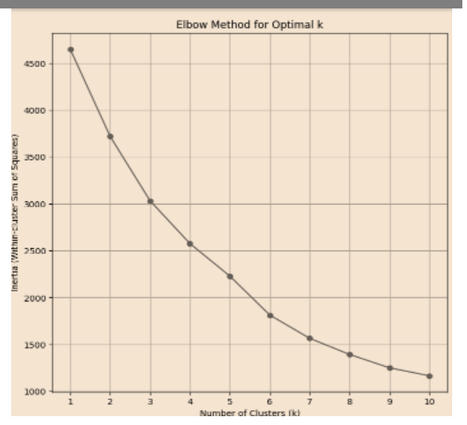


FEATURE ENGINEERING

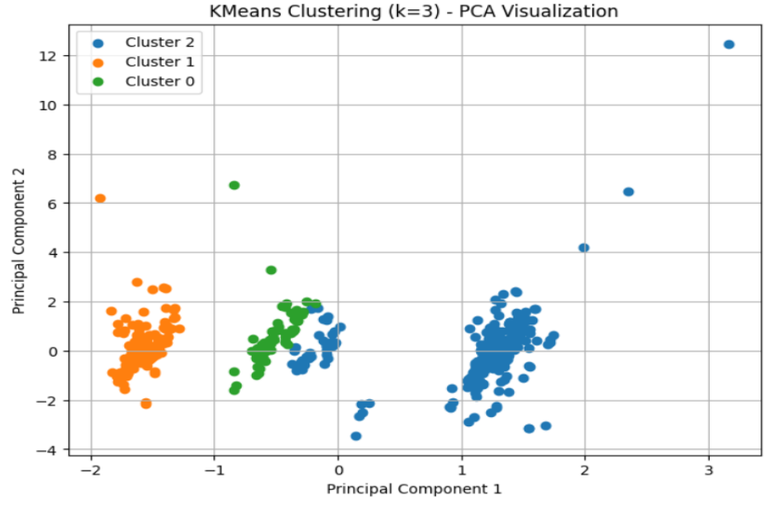


KMEANS CLUSTERING MODEL

- Elbow Method to find the optimal numbers of clusters
- K=3



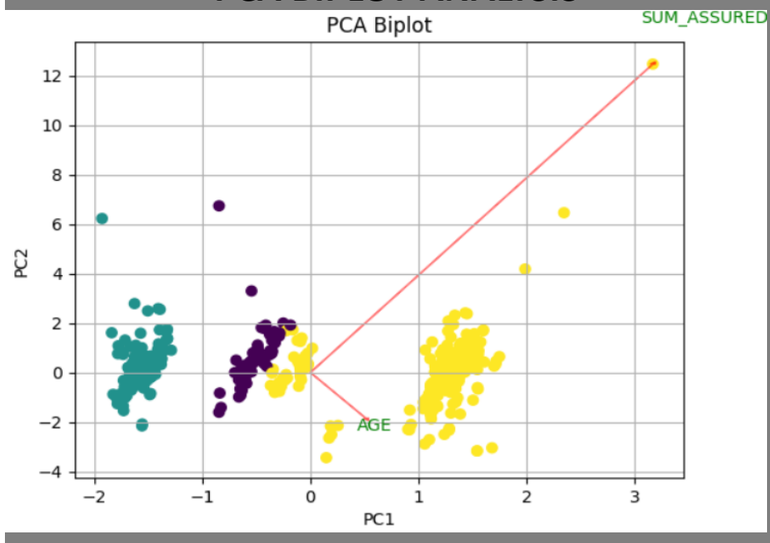
PRINCIPAL COMPONENT ANALYSIS




```
Explained Variance Ratio: [0.21508437 0.17046406]
Total Explained Variance: 0.38554842980359383
```

- The PCA graph shows 3 distinct clusters of the data.
- Explained variance ratio shows that 21% of the data features are explained by PC1 and 17% is explained by PC2.

PCA BIPLOT ANALYSIS



- Sum Assured has a strong positive correlation with PC1 and Age has a strong positive correlation with PC2

CONCLUSION

- Sum Assured and Age as shown above can be used in Segmenting customers according to various risk classes for accurate premium allocation.