Insurance Claim Analysis Demographics and Health

Welcome to our analysis of insurance claims data! Our goal is to identify patterns and clusters in the data, with a particular focus on age groups. Let's dive in!

by Praveen Karthik Arumugam

Linkedin - www.linkedin.com/in/pk7779

Github Portfolio - https://github.com/praveenkarthika/data_science.git

Problem Statement & Type

Problem statement

We aim to identify patterns in insurance claims data across different age groups.

Problem type

Using clustering techniques - KMeans and DBSCAN

Data Collection

1 Collection method

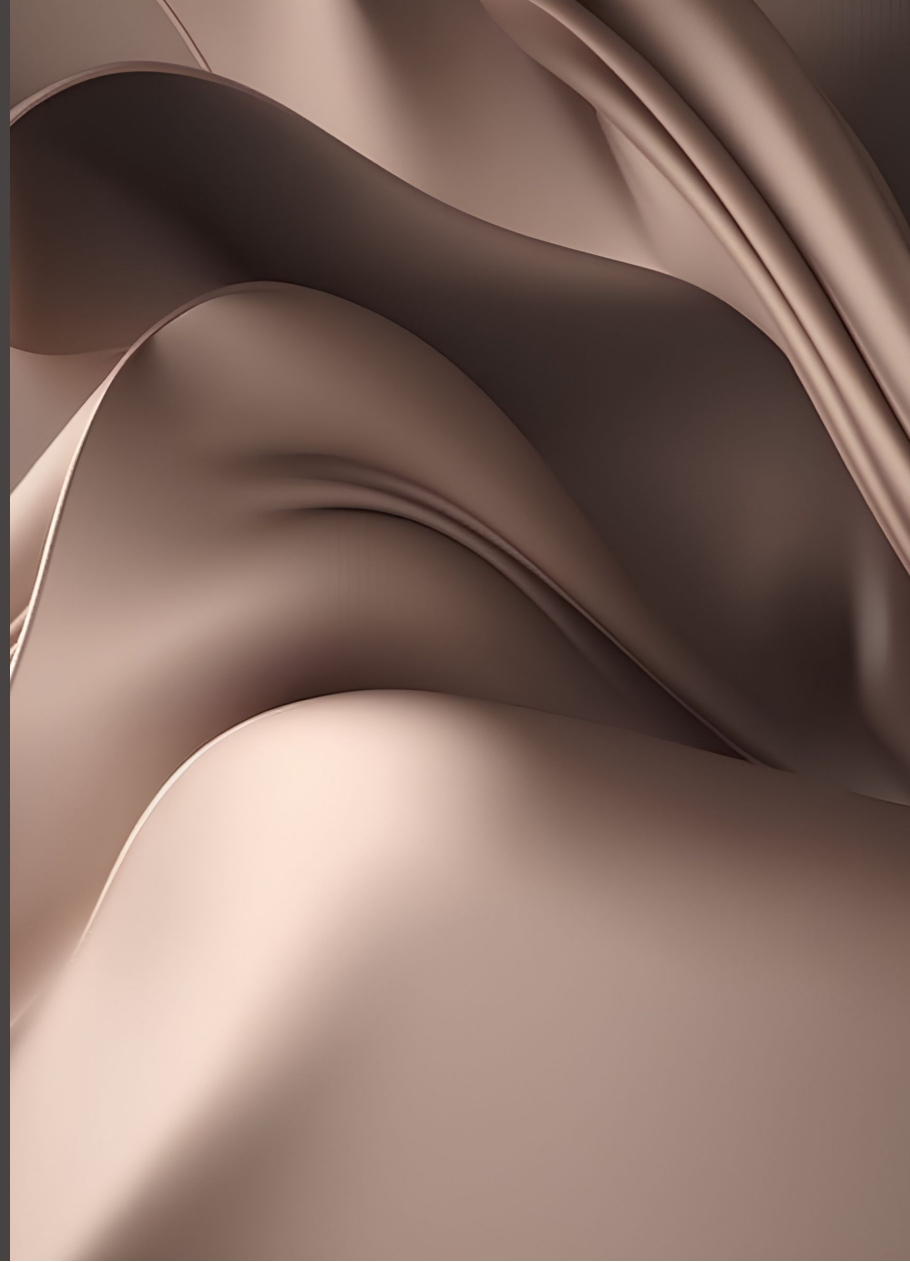
The data was collected via Kaggle - **Insurance Claim Analysis: Demographic and Health | Kaggle**

2 Dataset size

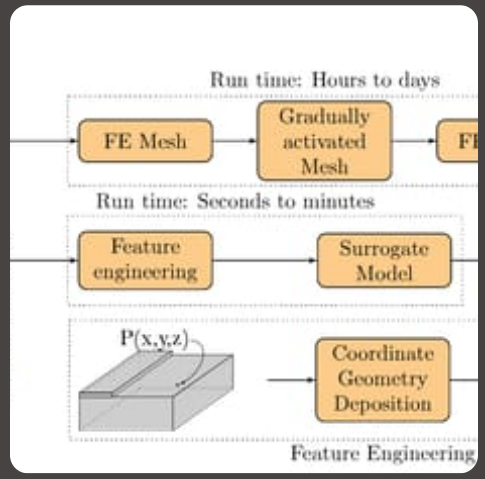
The dataset contains 1340 records

3 Features

Features include region, claim , ages, diabetic, smoker details

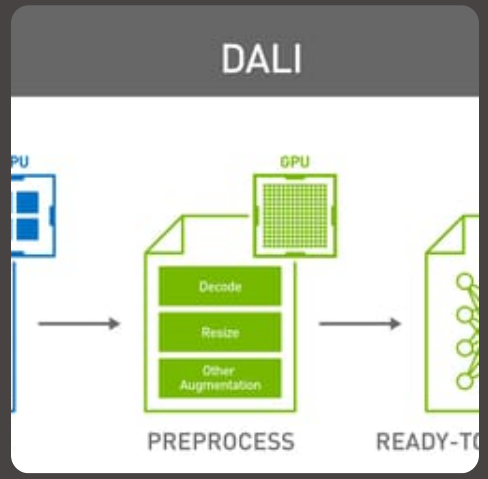


Exploratory Data Analysis



Feature Engineering

During feature engineering, we used LabelEncoder



Data Preprocessing

We handled NaN and null values by replacing them with Mean values

Data Visualization

To better understand the data, we created scatter plots

KMeans Clustering

1

Optimal K value

We found the optimal K value by using the elbow curve technique

2

Clustering data

We then applied KMeans clustering to our data

3

Results

A closer look at our results revealed 2 clusters

DBSCAN Clustering

Methodology

We used DBSCAN, a density-based clustering algorithm, to analyze our data. To find the optimal parameters, we performed a grid search and used silhouette scores.

Results

The best hyperparameters were epsilon and n_samples. After clustering the data, we discovered that optimal eps and n_sample was not appropriate

Conclusion

1 Key Findings

Most of the claim amounts were less than $15000 across all the age groups from 18 through 60