1. Create the BMI variable based on CDC definition1. Show your code or formula used, and the BMI values for the first three rows of data provided.

Code:

A screen shot of a computer code

Description automatically generated

Results:



1. Explore the data and report 3 key findings.

Seen in Figure 1, a new variable was created using ‘Age’ to separate Young (< 30 years old), Middle Aged (30 to 48 years old), and Old (>48 years old). We find that the Old age group pays a significantly higher premium compared to the other age groups, almost twice as much as the Young age group.

A graph of different colored squares

Description automatically generated

Figure 1. Age Group VS Premium Prices

In figure 2, we plotted non-binary variables in a correlation matrix. We find that Age is highly correlated to premium prices which reinforces our findings above, and also that Height, Weight, and BMI are uncorrelated to premium prices. Surprisingly, number of major surgeries (NumMajorSurgeries) also has a low correlation to premium prices. One would assume that an individual prone to suffering injuries might have a lifestyle revolving more risks, which should play a part in increasing their insurance premiums.

A screenshot of a graph

Description automatically generated

Figure 2. Heatmap between non-binary variables

In Figure 3A, arbitrarily setting eps to 0.5 and minimum samples as a conservative 5, we found a natural cluster across all age groups with BMI above 45. In Figure 3B, we notice that the cluster that exceeds a BMI score of 45 tend to pay significantly more in premium prices and are mainly seen to be men. However, their rate of critical illnesses appears to be lower than that of the rest of the population. This means that despite paying a higher premium due to their BMI alone, the cluster of clients above a BMI score of 45 seem to be extremely profitable – quite literally, a cash cow.

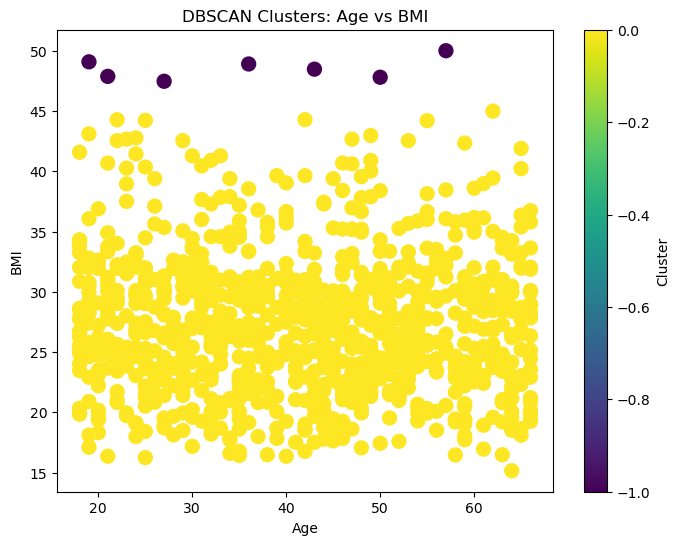


Figure 3A. DBScan Cluster between BMI and Age

A screenshot of a computer screen

Description automatically generated

Figure 3B. Cluster Summary

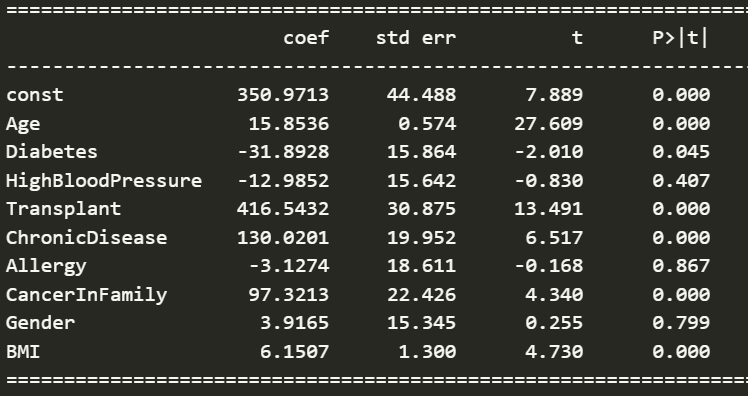
1. Using Linear Regression:
   1. Explain how you will select the “optimal” subset of X variables in your final linear regression model.
   2. Do a 70-30 train-test split and report on the testset RMSE. Explain the meaning of RMSE.
   3. Is BMI or Gender important in determining premium?
2. Having determined that there is multicollinearity between Height/Weight and BMI due to the calculations of BMI, as well as a low correlation between Premium and NumMajorSurgeries, we drop them as features for the regression. We do a preliminary regression using StatsModels’ OLS to explore the P-values of individual variables.  
     
   Here we find that Diabetes, HighBloodPressure, Allergy and Gender have a P-value >0.05, assuming a significance value of 0.1.  
     
   We conclude that the relevant variables for this model will be Age, Diabetes, Transplant, ChronicDisease, CancerInFamily, BMI.  
     
   In the finalized restricted model, we have:
3. Root Mean Squared Error (RMSE) is the square root of Mean Squared Error (MSE). It calculates the squared sum of residuals, summing the differences between actual and predicted values of the model in such a way that error values are positive and averaged across sample size.   
     
   Due to the squared exponent nature of MSE, outliers may heavily influence the accuracy metric, which is why RMSE is typically preferred and intuitive for interpretation by non-technical stakeholders.
4. 

Figure 5. Preliminary Unrestricted Model Summary (OLS)

In Figure 4, the unrestricted model identifies Gender as a statistically irrelevant variable  
 with a P – Value of 0.799 whilst identifying BMI as a statistically highly significant variable  
 with P – Value <0.001.