Question 3

Our project makes use of regression techniques to predict the response variable using a combination of the predictor variables. Regression analysis is a statistical process for estimating the relationship among variables [1]. Since we had to do predict a variable, we thought it was essential to understand the correlation between the various predictors. We used a scatter plot to analyze the relation among the significant variables, namely car\_value, age\_youngest & cost. These variables were selected as significant because they each had a p-value of less than 0.001 which was safely below the threshold of 5%.

Upon inspection of the scatter plot for record\_type = 1, it is observed that a majority of the data points are scattered around the region where the values of the 3 significant variables are:

1. car\_value between e and f
2. age\_youngest uniformly distributed between 30 and 55 years
3. cost between 625 and 700

car\_value e and f on a scale of c to I indicate that the customer’s car was moderately priced when purchased; cost between 625-700 on a scale of 400 to 750 indicate that the cost of the insurance was quite high. Most of the people purchasing the insurance were people between the ages of 30 and 55 years, which could possibly indicate that people who were of considerable working age and who might be married are most likely to purchase the insurance.