Objective

The objective of this task is to find out the key factors that affect the cost of healthcare by analyzing data obtained from patients hospitalized for a certain condition.

Data

We have 4 tables, which contain patient medical as well as financial data. It is organised as:

Bill\_amount.csv – Costs of each patient visit

Bill\_id.csv – Maps each bill to the appropriate patient

clinical\_data.csv – Clinical data collected for each patient visit

demographics.csv – Patient demographic information.

These may be combined using the bill id (unique per visit), patient id (unique for patient) and the date of admission (unique per patient and visit).

On joining the given tables, we have a set of 13600 unique patient visits, with a corresponding cost of healthcare for each visit (which is the response we want to predict).

We have 28 factors which are our predictors (excluding keys like patient id, bill Id and date of admission).   
These include clinical data such as height, weight etc, demographic data such as race, gender etc as well as some patient clinical history details, hospital procedure records, patient symptoms summary and some lab tests.

Data preparation

Among the predictors, we have 5 string columns while the rest are numeric. The string columns are medical\_history\_3, gender, race, residential status and date of birth.

Gender, race and residential status have limited possible outcomes and can be converted into numeric factors by using dummy variables for each possible outcome. Misspellings and missing data also need to be handled.

Medical\_history\_3 seems to be a Boolean factor but with Yes/No that can be replaced to make 0/1.

Date of birth may not be very useful when considered by itself but we can use this information to calculate the age at the time of admission.

Checking for missing data, ‘medical\_history\_5’ and ‘medical\_history\_2’ are 0/1 columns that have significant number of missing data, and hence we can either impute this data or drop these predictors. Given the large percent of missing data, I decided to consider missing observations as a separate category and marked them as such.

Feature engineering

Based on rudimentary knowledge of some of the predictors, we can create some composite features. These can help increase our insight on the significant factors.

For eg: weight and height can be combined to form a BMI.

Number of symptoms/Number of pre-operation medications can be summed up together as more symptoms/more medication may mean higher costs of treatment.

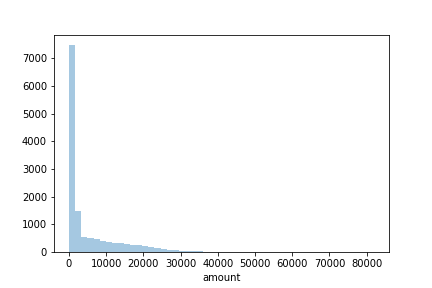
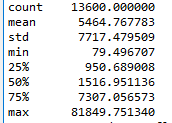
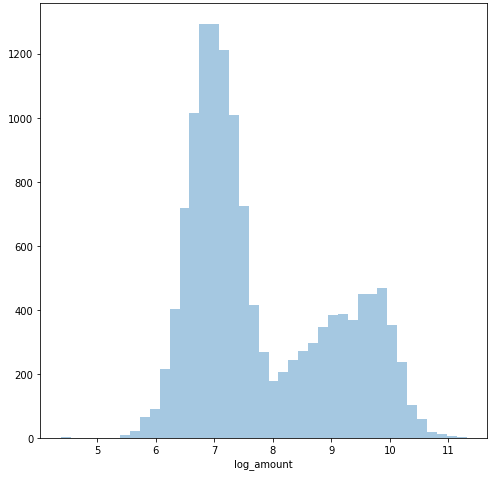
Total time spent in treatment can be calculated. If a patient is hospitalized for more time costs are likely higher

Data analysis

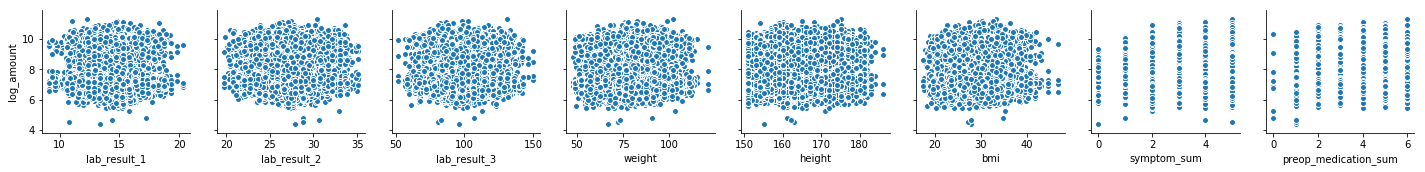
After cleaning the data, we begin to look at the response and the factors.

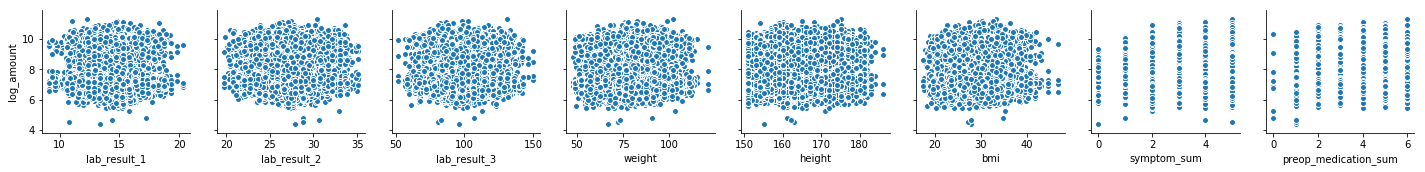
Looking at the response, we can see that there is a long tail of amounts and the distribution is highly skewed. Thus, it may be a good idea to use a log of the amount as our response variable instead.

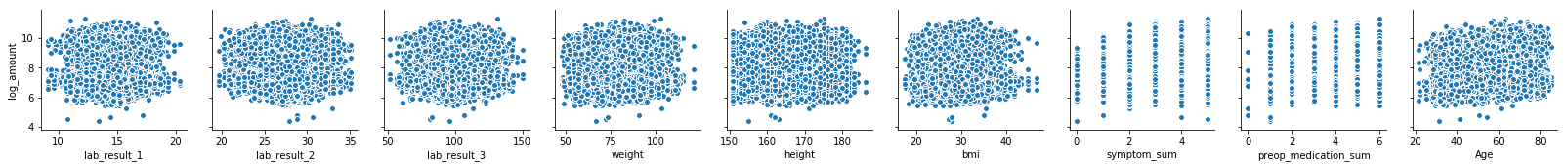
It seems our response is now bimodal with 2 distributions, and we need to analyse where this may come from.



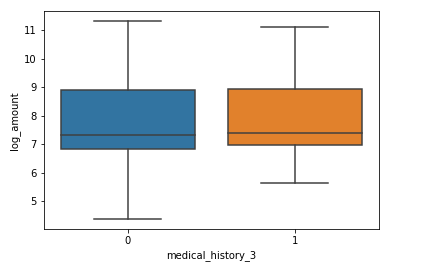
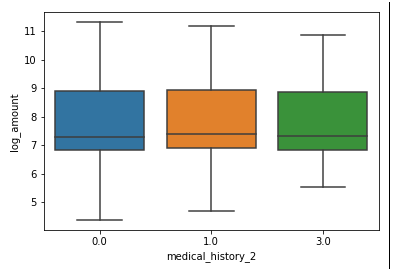
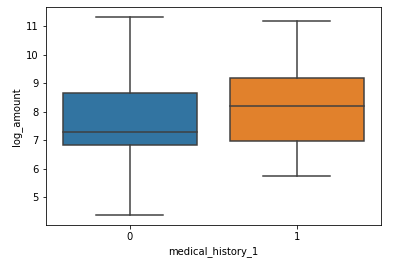
Looking at the univariate linear correlations for the response against each of the continuous value predictors:

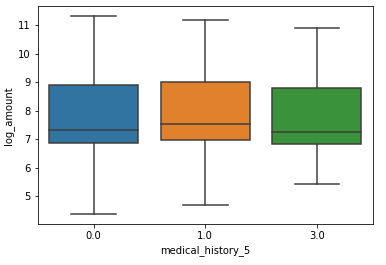
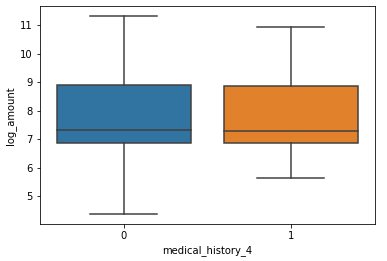
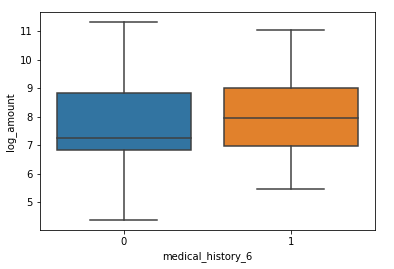


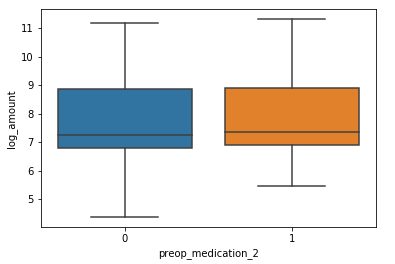
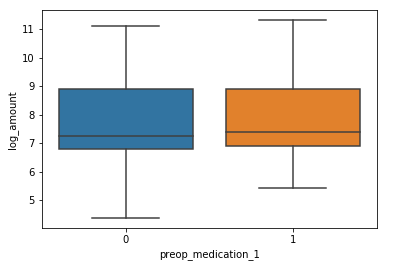
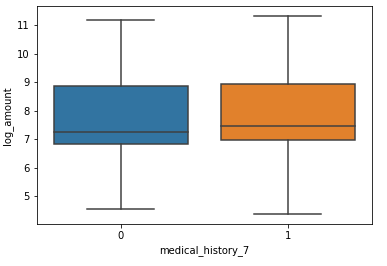


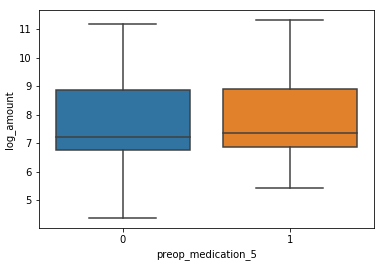
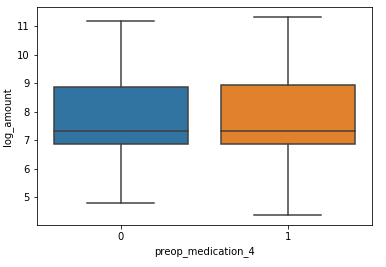
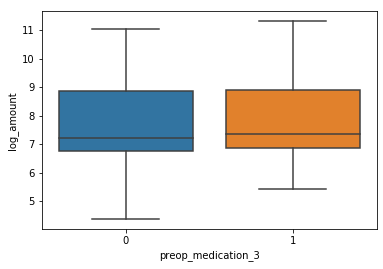


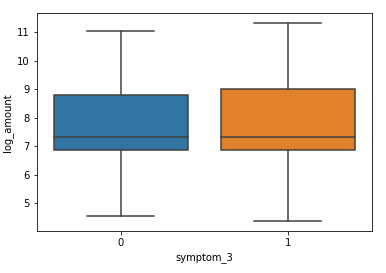
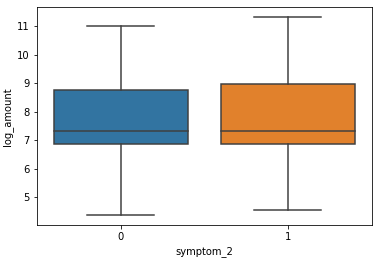
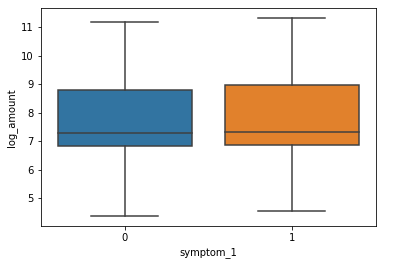
Looking at the categorical factors:

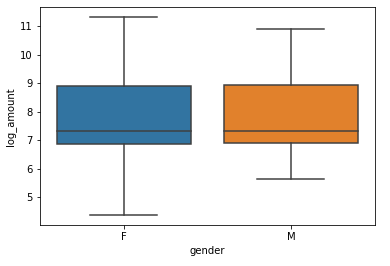
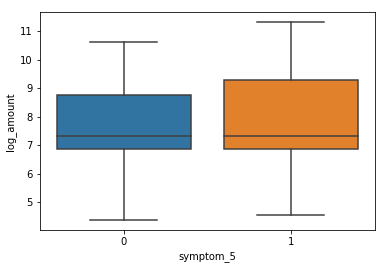
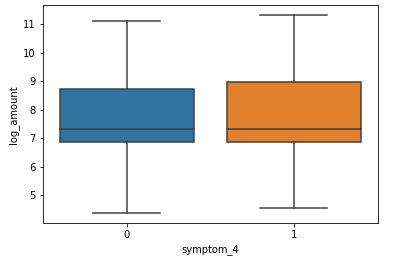


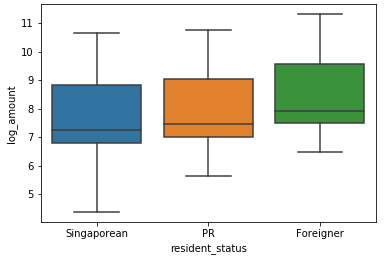
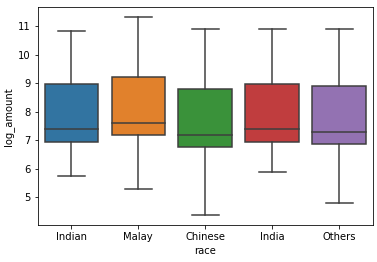
 











Based on these plots we are not quite able to judge the strongest predictors. Some highlights, though are:

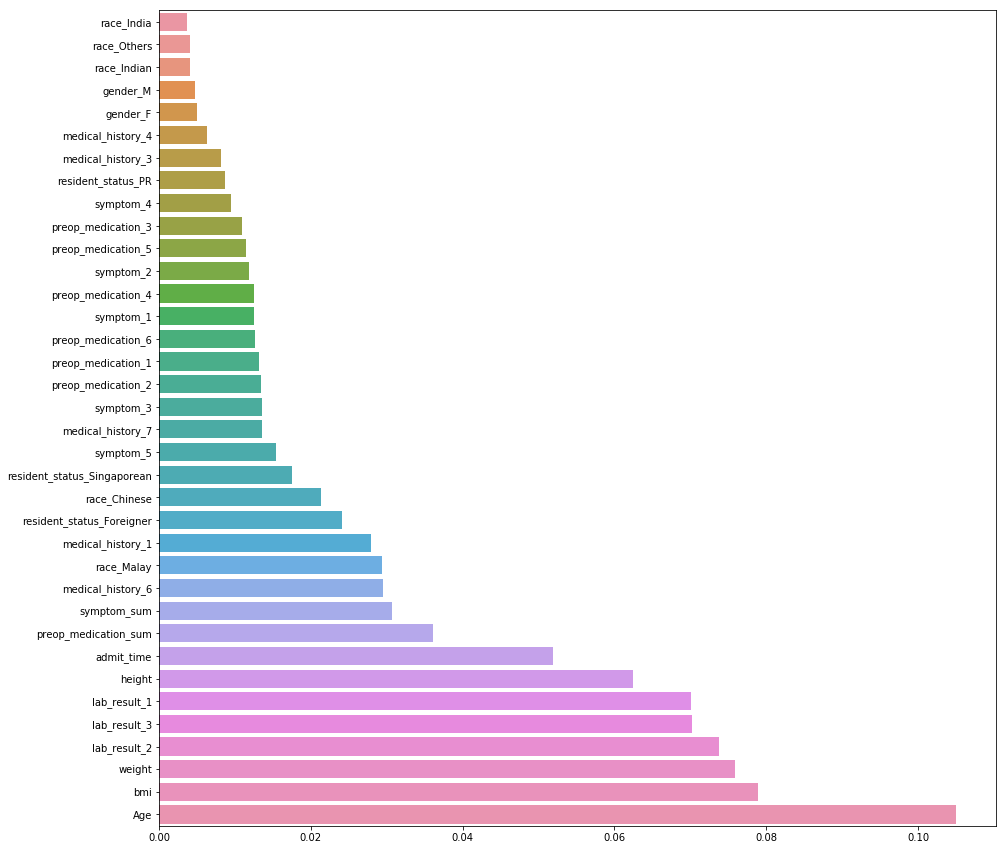
* Age seems to be correlated positively to the cost of healthcare, which may be expected as older patients may need more treatments.
* Total number of symptoms is correlated but the effect stagnates as number of symptoms increases.
* The medical\_history\_1 and medical\_history\_6 are important predictors.
* Females have higher variance compared to men, in the cost of treatment received.
* Race and Resident status also seem to affect the total cost.

In order to better identify the key features we use a univariate feature identification test and a random forest based regressor to identify and rank important features.

Based on the univariate analysis- following factors are selected;

‘medical\_history\_1', 'medical\_history\_6', 'symptom\_5', 'Age’, 'symptom\_sum', 'preop\_medication\_sum', 'race\_Chinese', 'race\_Malay', 'resident\_status\_Foreigner', 'resident\_status\_Singaporean'

Random forest based classifier gives us the following features:



Based on this, the most important factor is age, followed by bmi , lab results and admitted time. Now we can look at each of the features individually, and work further on creating features that can explain the data better.