Holmusk Assignment Report Draft 1

# Introduction

# Data Preprocessing & Feature Engineering

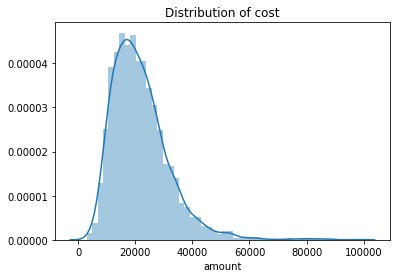
## Preprocessing

Before

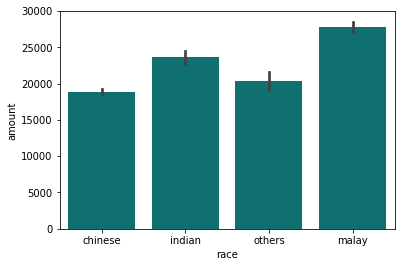
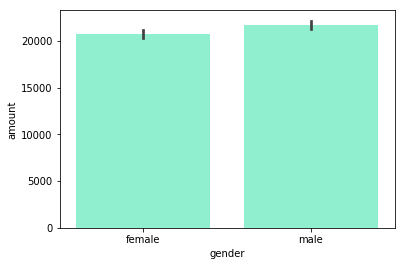
## Feature Engineering

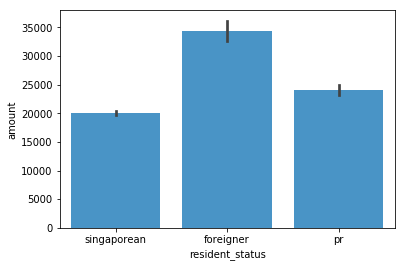
## Exploratory Data Analysis (Univariate & Multivariate)

1. Distribution of cost

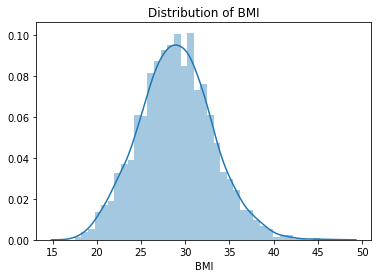


1. Distribution between gender

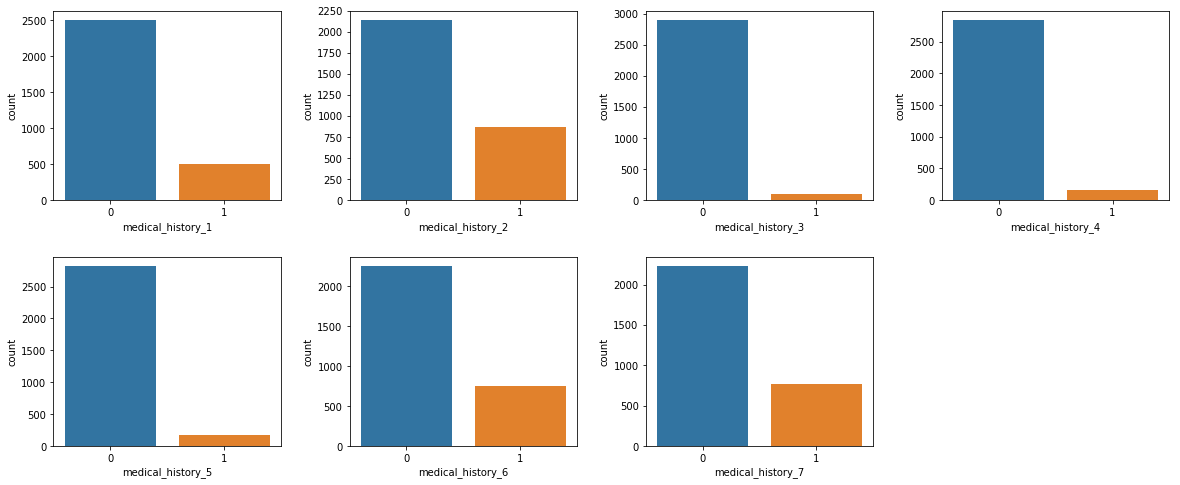




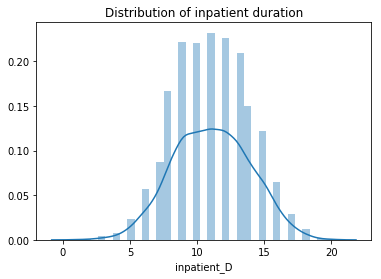
1. Distribution between races
2. Distribution between age
3. Distribution between BMI

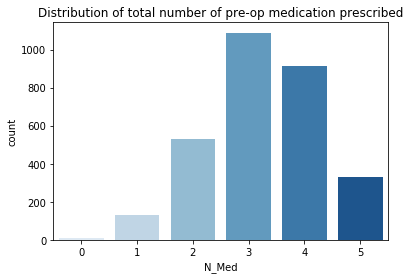


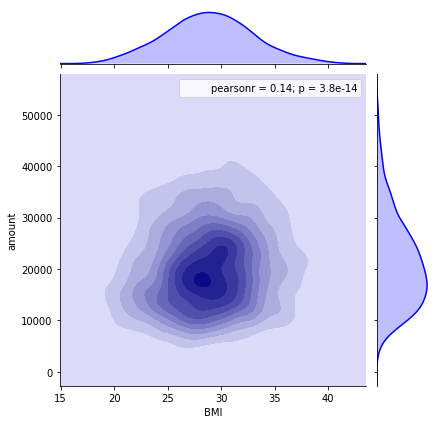
1. Distribution of lab tests
2. Distribution of medical history availability

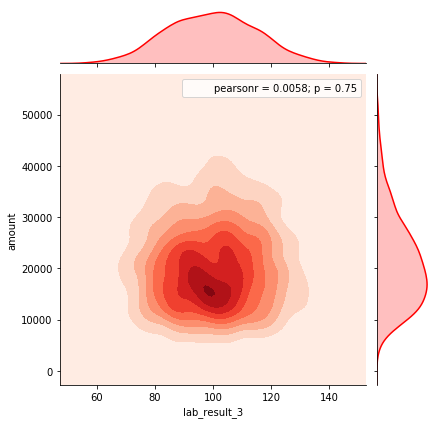
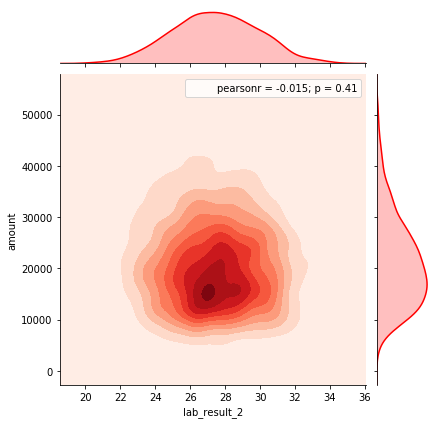
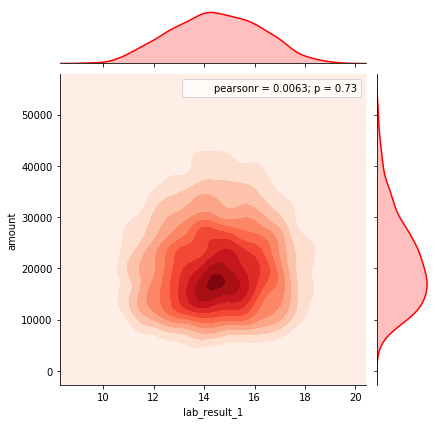
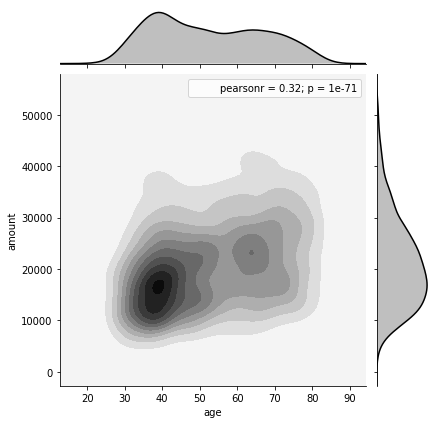


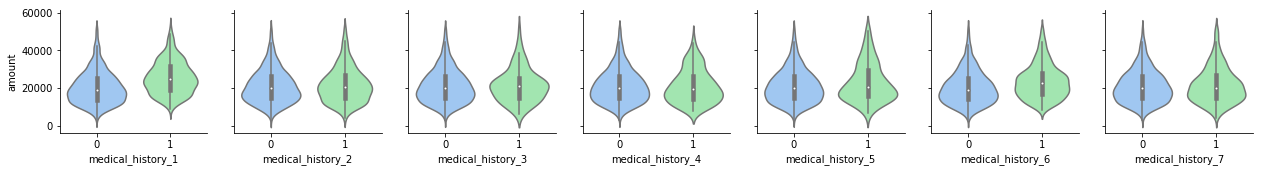
1. Distribution of inpatient duration

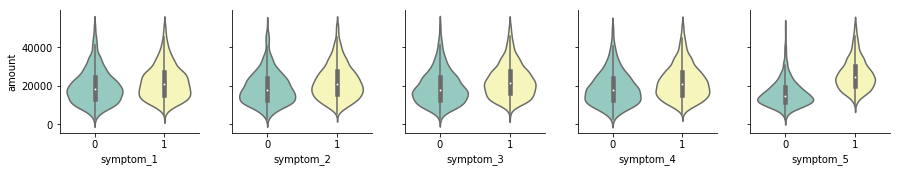


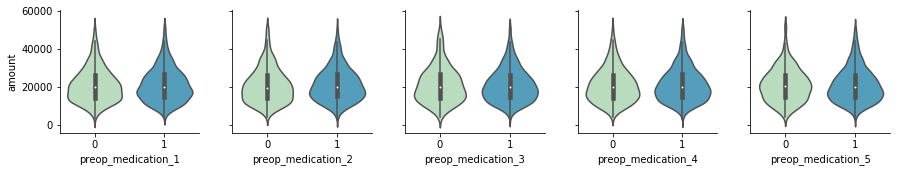
1. Distribution of pre-operation medication
2. 
3. Distribution of total number of specific symptoms exhibited
4. 

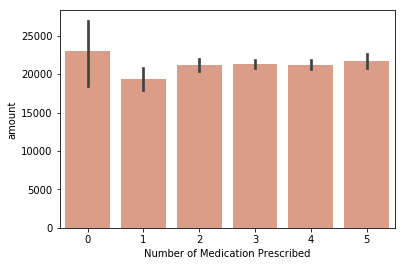
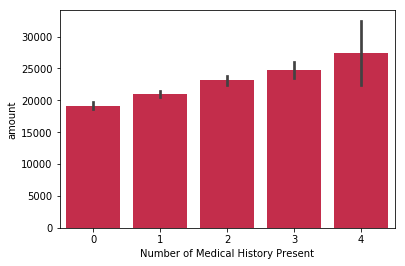


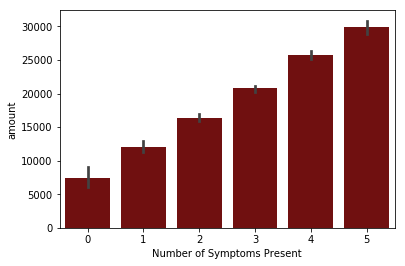


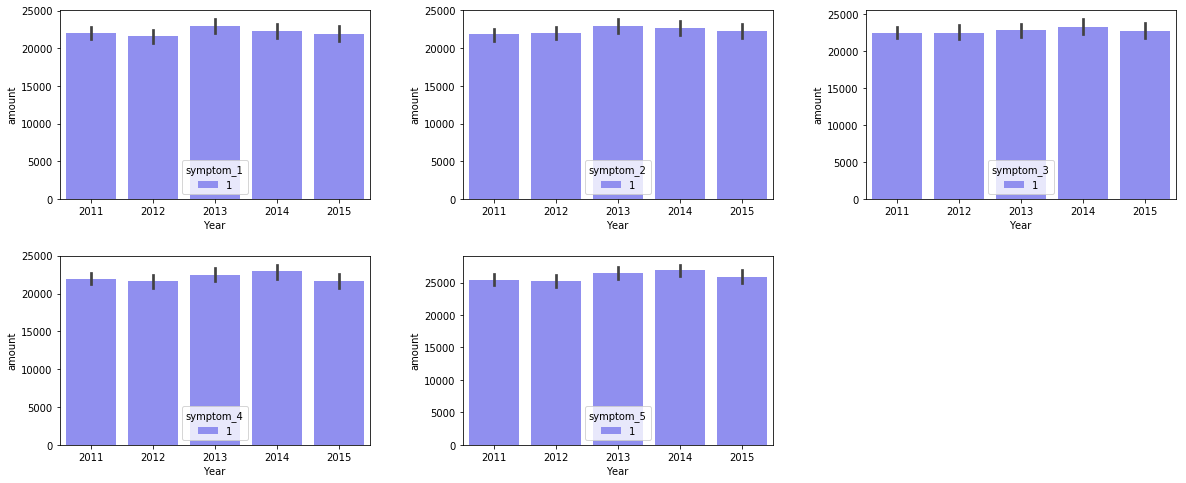
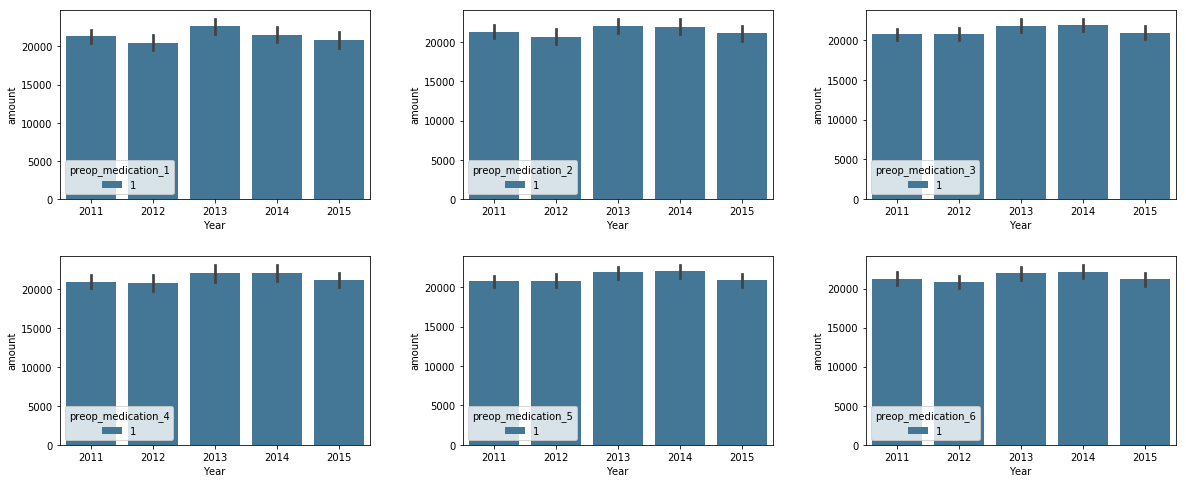
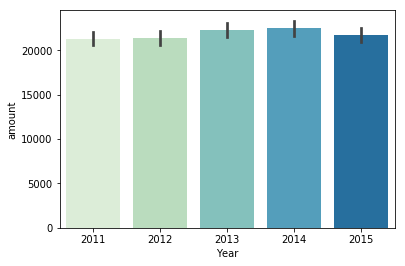


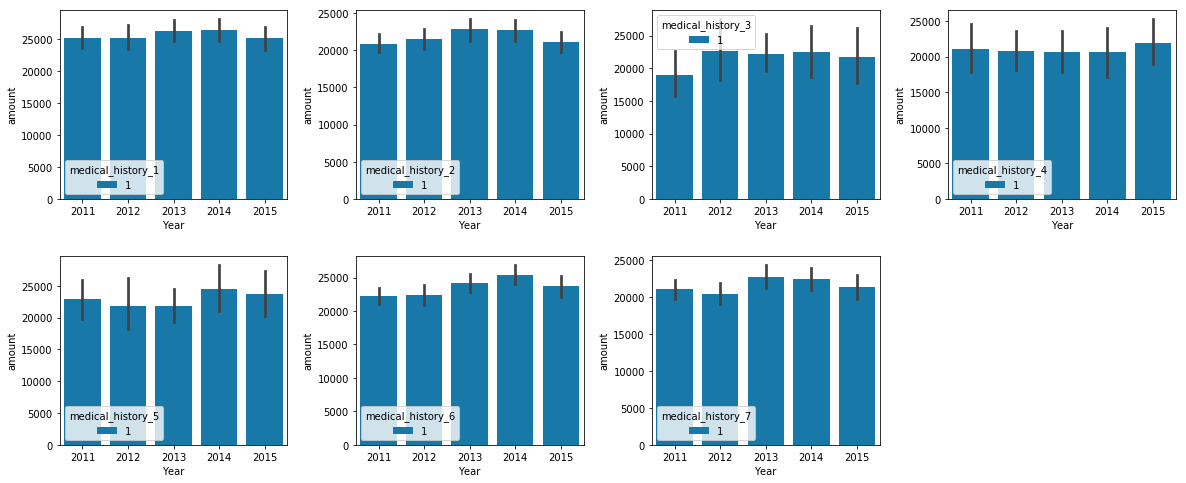


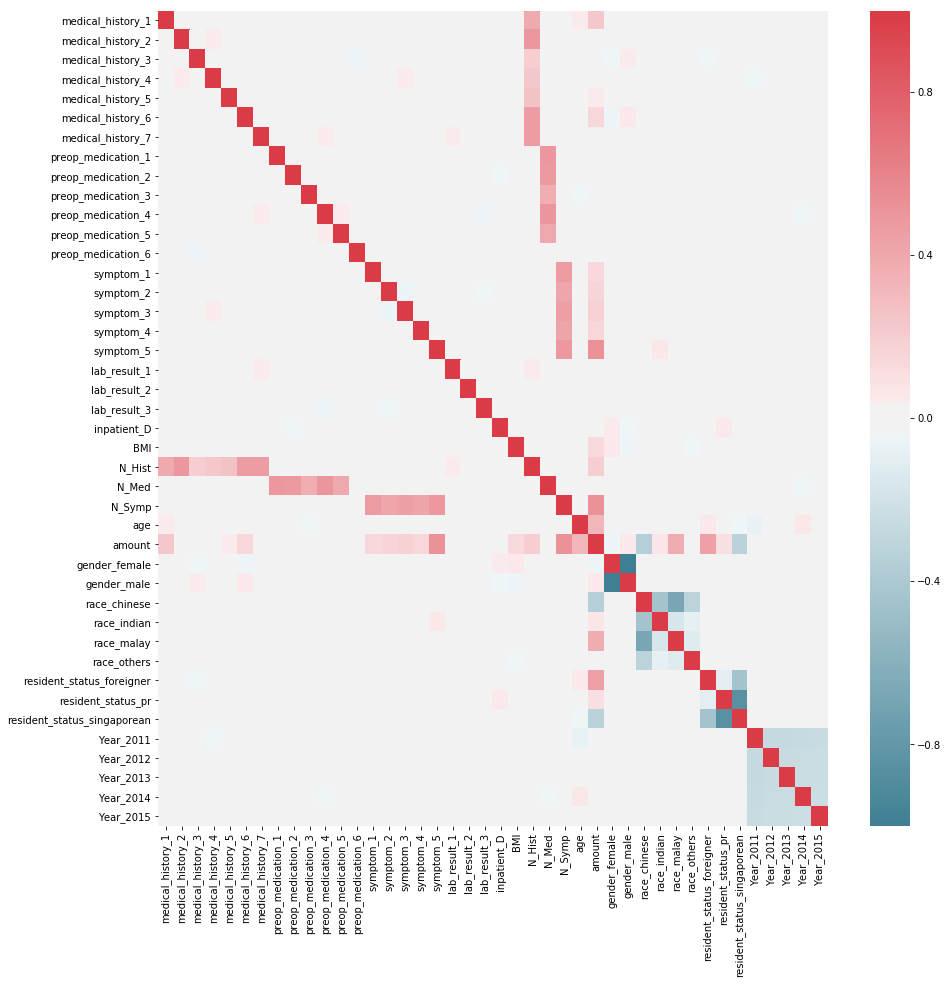












### Method

### Random Forest and Extra Trees Regression Model

Random Forest(RF) algorithms are a form of supervised learning models developed to overcome the overfitting problem that decision trees algorithm faces, which lead to high variance and low bias. RF algorithms build multiple decision trees by training different trees on different on different parts of the training set via bootstrapping. Additionally, at every node of the decision trees, a randomly selected sample of features from the dataset are chosen. A condition is applied to each feature within this sample, designed to split the dataset into two an ensure that similar response values are grouped into the same set. The feature which best meets the condition at that node is then used to split the trees and the process repeated. These multiple decision trees are then averaged to build the predictive model, which can be used for classification and regression.

The ExtraTrees (Extremely Random Trees) algorithm introduces a higher level of randomness via two main changes to the RF algorithm. First, instead of bootstrapping the training dataset, the entire dataset is used during training. In addition, instead of choosing the feature which best meets the condition set at the node, the feature is chosen at random. Removing the bootstrapping reduces bias in the training dataset, while randomizing the cut-point leads to variance reduction in the ExtraTrees forest model.

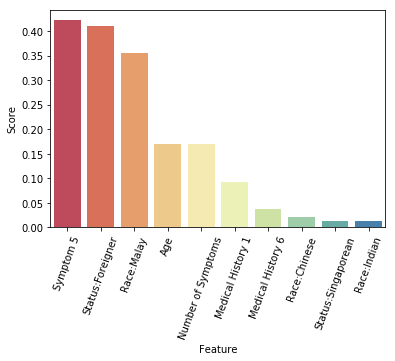
### Feature selection strategy

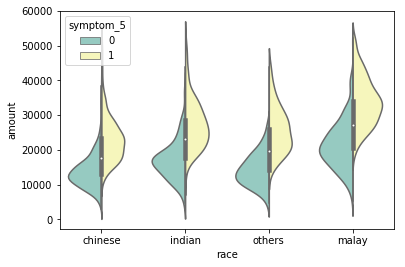
One strong advantage of RF models are their feature selection capabilities through feature importance computation for each predictor feature. At each node of the many decision trees, the measure used to determine the locally optimal condition is mean squared error(MSE). This measure can be computed even if it is not used in the cut-point selection. Hence, it is possible to compute the MSE decrease for each feature in each tree, average them across the forest and rank them, allowing us to identify features which contribute most to the regression model.

However, one limitation of feature selection of RF models is that it is biased towards preferring features with a larger number of categories. The second limitation is potential lower reported importance for correlated features. This limitation arises with correlated features because if one correlated feature has already been selected at an earlier node, MSE reduction of the other correlated feature contributes has already been removed.

To counter these limitations, one method is to compute the mean decrease in accuracy by directly measuring the impact of each feature on the model accuracy. This is done by permuting the values for each feature and measuring the mean decrease in accuracy. Relatively unimportant variables will show small decreases in accuracy, and vice-versa. The sklearn API does not currently support this capability for RandomForestRegressors and ExtraTreesRegressor so this feature was implemented inside the code.

# Results





# Discussion

Statistical Significance vs Practical Significance

Existing Literature

# Conclusion

# Appendix