**Holmusk Assignment Report**

# **Data Preprocessing & Feature Engineering**

The datasets assigned were first visually examined using Microsoft Excel to for an initial examination. The csv files were then opened on a Jupyter Notebook as a pandas (pd) DataFrame. DataFrame attributes ease the understanding and exploratory data analysis (EDA) of data structures.

## **Data Cleaning**

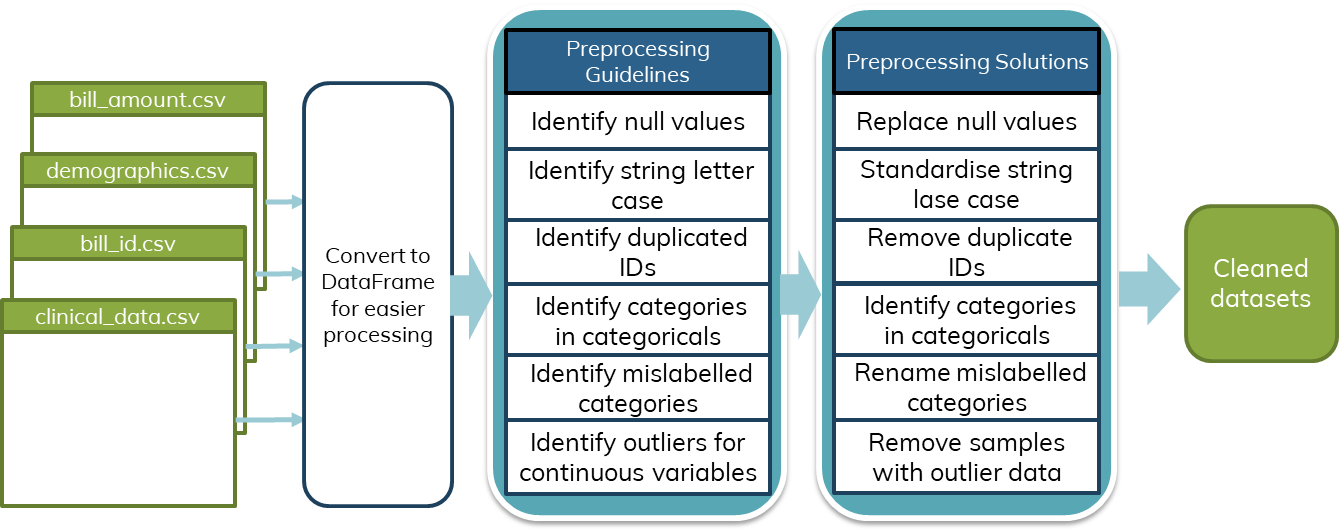
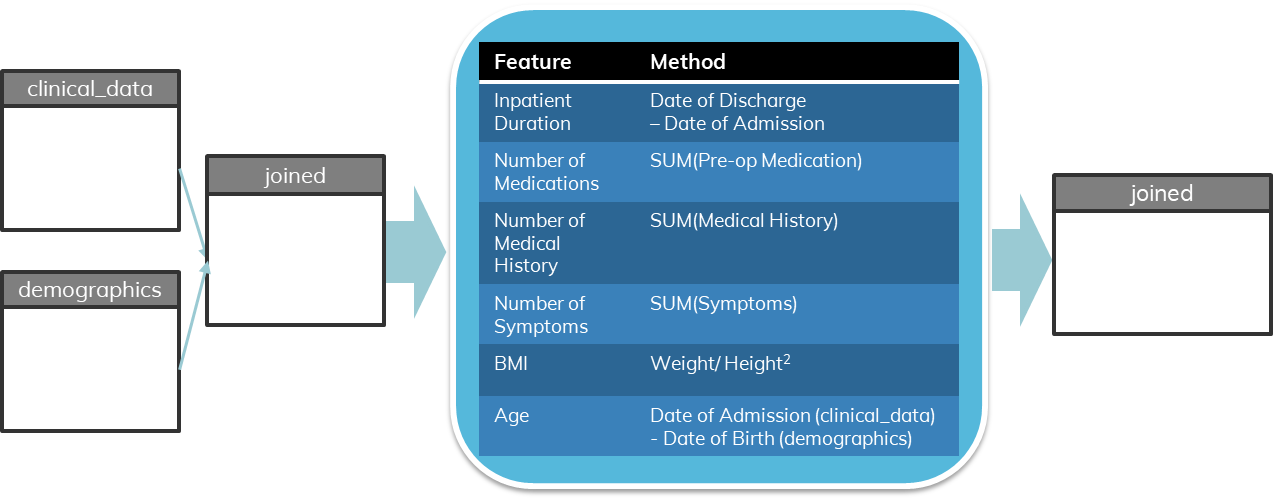


Figure 1. Data cleaning pipeline diagram

The data pre-processing followed a list of guidelines which can be visualised in Fig 1. DataFrame attributes were used to quickly identify null values and the different categories in categorical features. Additionally, the letter case for categories in categorical features were standardised to lower case for easier processing and analysis. Mislabelled categories include features whose categories may be represented in different string formats (E.g. India vs Indian under the race column). Mislabelled categories were easily identified using grouping attributes and replaced with the correct category names.

## **Feature Engineering & Dataset Simplification**



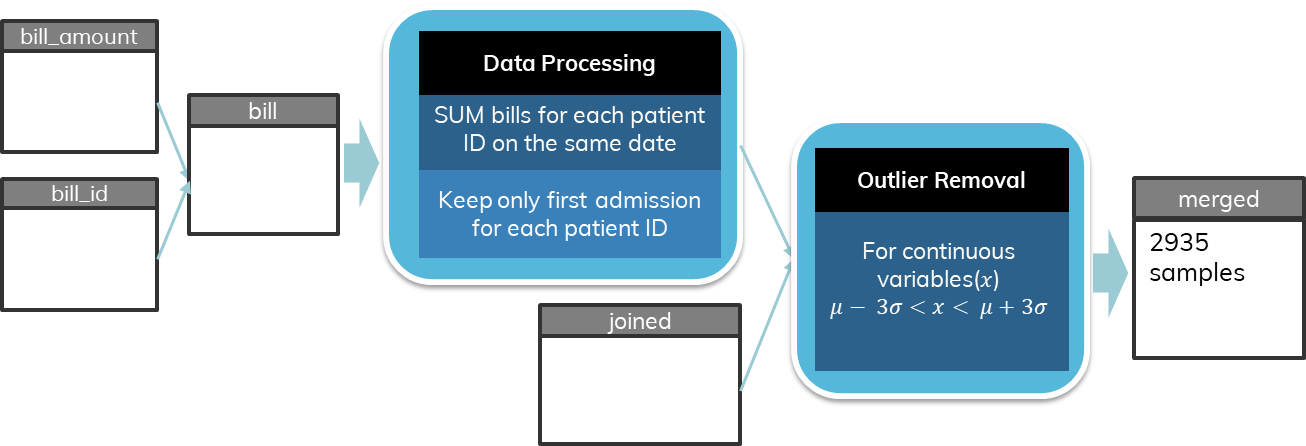


Figure 2. Feature engineering and datasets merging pipeline.

The clinical\_data and demographics dataset were first merged, and feature engineering conducted on the joined dataset. From visual examination of the features in the two datasets, prior understanding and knowledge of healthcare lead to the realisation that there could be engineered features which would more likely be a potential driver of cost of care. These features are i) inpatient duration, ii) number of medications, ii) number of medical history, iv) number of symptoms exhibited, v) BMI and vi) age.

Inpatient duration could be derived from the dataset by subtracting the date of admission from the date of discharge, which can be easily done on a pd DataFrame. The three categorical variables in the clinical\_data dataset; preop\_medication, medical\_history and symptom are in the format of an indicator matrix. As such, it would be possible to determine the total number of preop medication prescribed to a patient, the medical history or symptom exhibited by aggregating the relevant columns. Age is another feature that can be derived from the datasets, although the date of admission and date of birth are located within two different datasets. Hence, the two datasets were merged first before feature engineering.

Visual analysis of the bill\_amount.csv file showed that there were multiple bill IDs associated per patient ID. Grouping of samples over patient ID and date of admission showed that there were 4 bill ID per date of admission per patient. As such, bill amount within the same day and patient ID were grouped together.

In addition, multiple patient IDs were present within the dataset. 779 samples out of 3400 unique patient ID – date of admission samples were found to have a repeated patient ID. Out of this 779 samples, there were 379 unique patient ID, indicating that 400 samples were patients readmitted once or more into the system. This will be further discussed in the Limitation section. The dataset was simplified by keeping the first admission for each unique patient IDs and removing all subsequent admissions. This lead to a sample loss of 379 which constituted 11.147% of the total sample. However, due to the relatively large sample size after sample size reduction (N = 3000), this simplification was carried out.

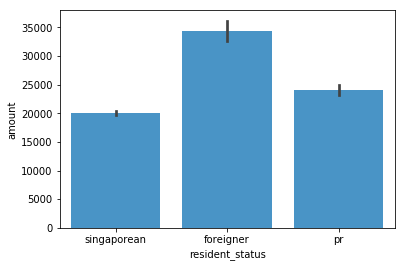
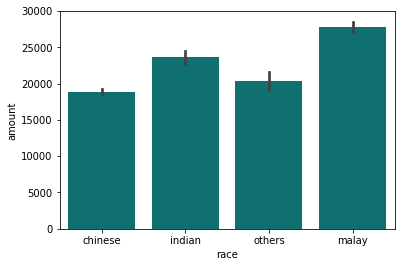
The merged bill\_id and bill\_amount datasets were merged on the bill IDs, and then merged with the joined dataset from clinical\_data and demographic on patient Outlier removal was performed on continuous variables in the resulting dataset, retaining samples which 3 standard deviations away from the mean.

# **Exploratory Data Analysis (Univariate, Bivariate)**

## **Univariate Analysis**

Univariate analysis was conducted on the categorical variables, preop\_medication, medical\_history and symptom. The distribution of the cost of healthcare was examined as well. However, no insight relevant to the cost of healthcare was obtained from univariate analysis. The figures generated can examined in the Appendix.

## **Bivariate Analysis: Descriptive Statistics of Features against Cost**



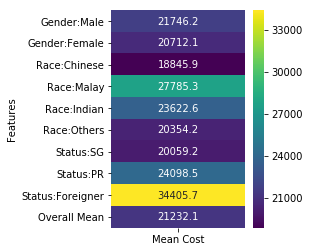
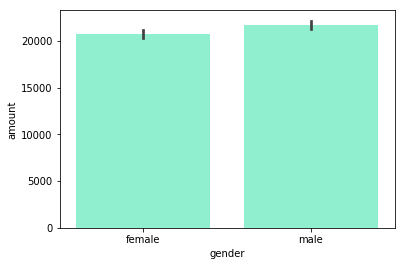
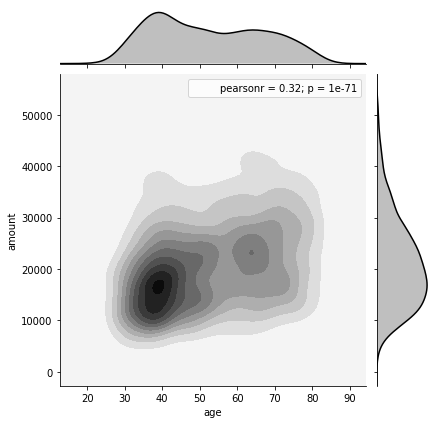
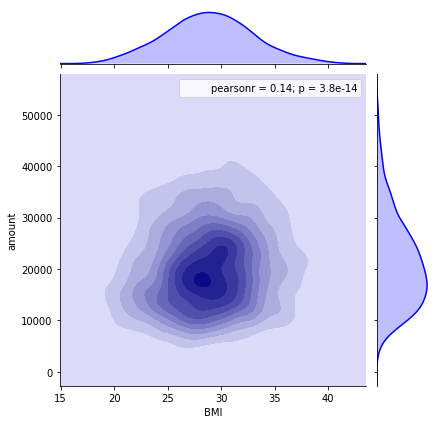


Figure 3. Descriptive statistics of categorical variables (race, resident status, gender) against cost

Bar plots of the categorical variables in demographics were plotted against cost. The bar plots generated show interesting results. The mean cost of healthcare for the patients of Malay racial status was the highest amongst the 4 categories, followed by patients of Indian, Chinese and Others racial status. The mean cost of patients for both Malay and Indian races (27785, 23622) were higher than the overall mean (21232). The mean cost for foreigners (34405) were higher than both Singaporeans and PRs(20059, 24098) . Both foreigners and PRs have mean cost of care higher than the overall mean. Gender differences do not seem to contribute to any differences in cost of care, with both mean cost being similar to the overall mean.



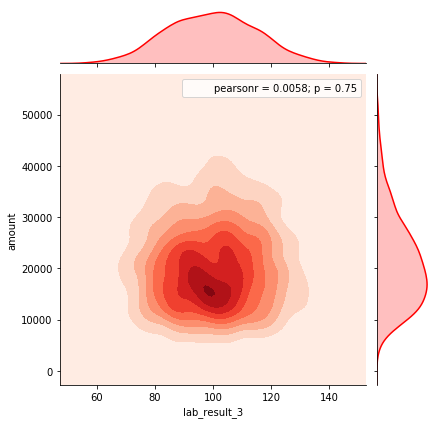
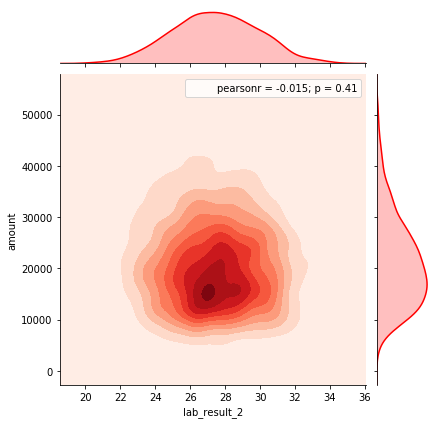
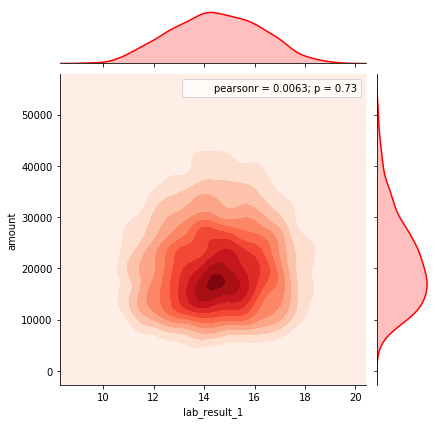
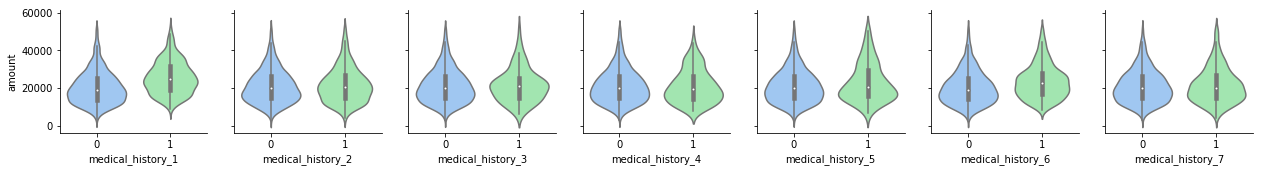
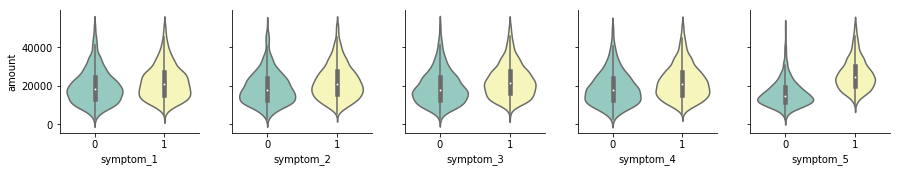


Figure 4. Kdeplots of continuous variables (BMI, inpatient duration, age, lab results 1, 2,3) against cost

For continuous variables, kernel density estimate (kdeplots) were plotted out. Kdeplots allow visualisation of the distribution of the variables with respect to the response variable. This would allow readers to view any potential linear relationship as well as the distribution density. Out of the 6 continuous variables, Age shows the highest absolute value of the Pearson’s r coefficient, at 0.32. The kdeplot of age also shows a slightly positive linear relationship between age and time. In comparison, the kdeplot of inpatient duration, lab result 1, 2 and 3 do not show any linear relationship with cost. This is also reflected in the near-zero Pearson’s coefficient for these features.





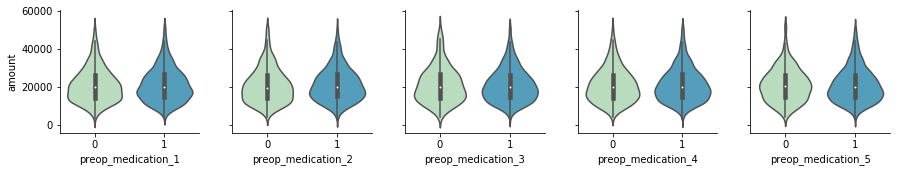


Figure 5. Violin plots of categorical variables in indicator matrices (pre-op medication, medical history, symptom) against cost

The three categorical variables already present in an indicator matrix were plotted using violinplots. A violinplot format was used to visualize the distribution of these variables against cost because it allows for side-by-side comparison of the distribution mean (the ‘fattest’ part of the violin) when the variable is absent (0) or present (1). In Figure 5 above, the variables which show a higher mean cost have been boxed in blue for ease of observation. Thus, medical\_history\_1 and symptom\_5 are potentially key drivers of the cost of care.

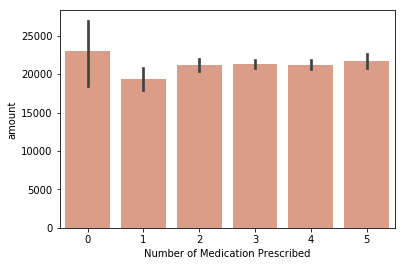
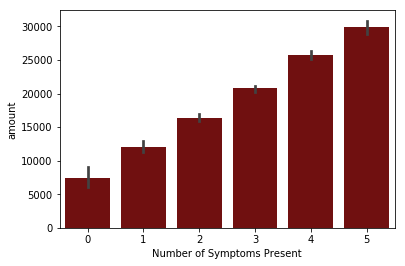
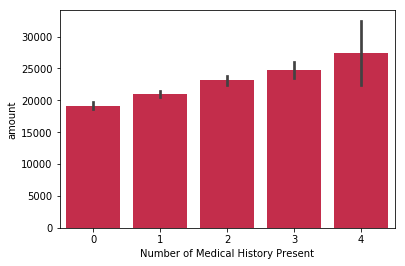


Figure 6. Bar plots of engineered features (Number of medical history, symptoms or pre-op medications prescribed) against cost

Although the engineered features are continuous in nature, the small range of the absolute values within these features allow for effective visualisation as a bar plot. The bagplots shown in Figure 6 above show a linear trend in increasing cost for both an increasing number of medical history present as well as an increasing number of symptoms present. Interestingly, there is a lack of linear relationship between the number of preop medications prescribed and the cost of care.

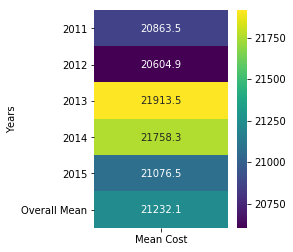
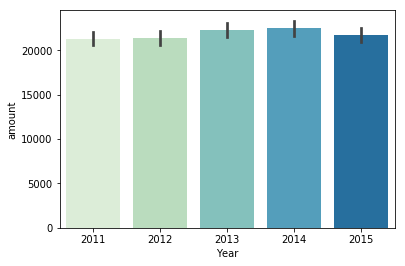


Figure 7. Descriptive statistics of year against cost.

One factor that I investigated was the effects of any year-on-year increase in cost. The mean cost of care from 2011 to 2015, which is the range of years for the samples in the dataset increased from 2012 to 2013 but decreased in 2015. In addition, the 2013 mean was slightly higher than the dataset’s overall mean. This implies that the year could potentially be a driver in cost of care.

## **Correlationship Analysis**

Correlationship between the features and amount was visualised using a heatmap. Categorical features which were not in an indicator matrix were converted into an indicator matrix using one-hot encoding.

The correlation matrix is visualised as a heatmap in the figure below. The heatmap shows that there is little correlation between the predictor features within the dataset. Collinearity is observed between the number of history and medical history 1-5, the number of medication and preop medication 1-7, and the number of symptoms with symptom 1-5 because the former feature is an aggregation of the latter features. Features derived from one-hot encoding categorical features show a negative correlation with each other due their antagonistic nature.

If we focus on the amount row or column, we can observe that a few features are positively correlated with this response variable. These are:

1. medical\_history\_1, 5 and 6
2. symptom\_1 to symptom\_5
3. BMI
4. N\_hist (Number of medical history)
5. N\_symp (Number of symptoms)
6. Age
7. Indian and Malay races
8. Foreigner Status

Hence, we can expect these features to be some of the key drivers of cost of care.

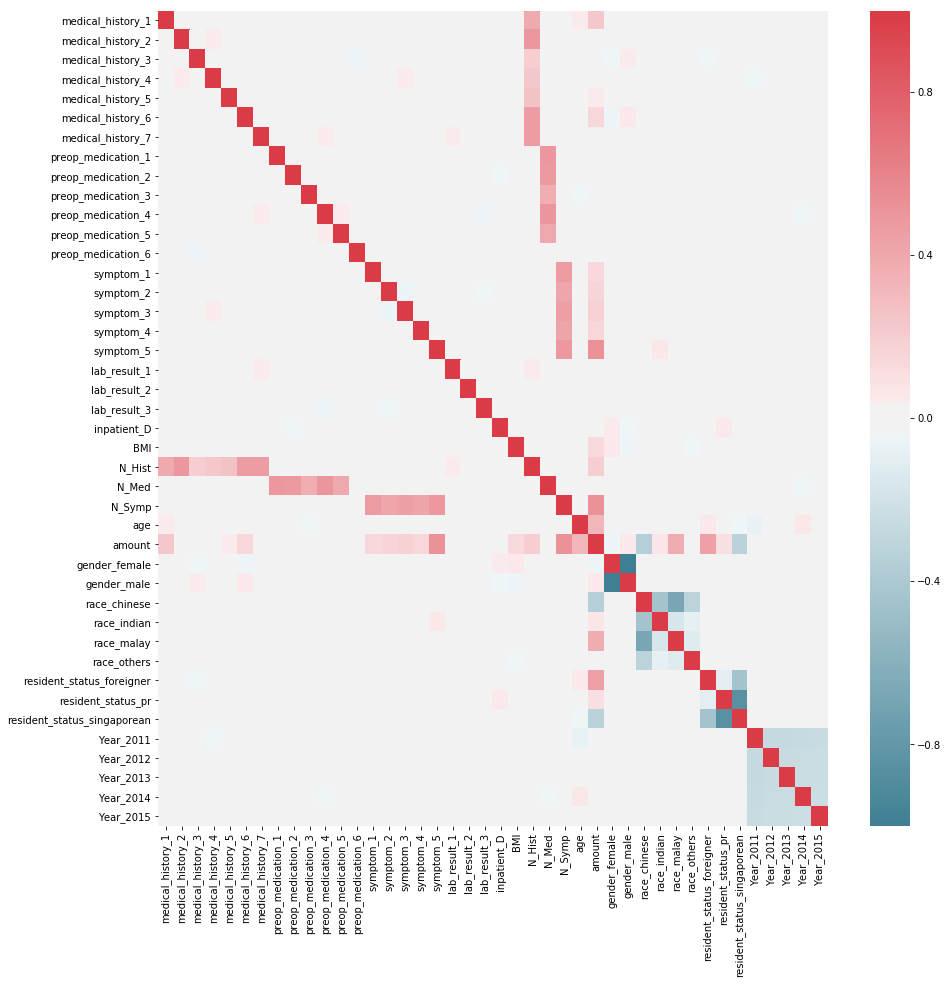


Figure 8. Heatmap visualisation of correlation matrix

# **Regression Modelling**

## **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.[1] 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 via r2 score. This is done by permuting the values for each feature and measuring the mean decrease in r2 score. Relatively unimportant variables will show small decreases in r2 score, and vice-versa. The sklearn API does not currently support this capability for RandomForestRegressors and ExtraTreesRegressor so this feature was implemented inside the code.

## **Model Training**

Before training of the Extra Trees Regressor model, continuous numerical variables were normalized.

The scikit-learn API was used to build the model. [2] In addition, the feature selection ranking method described in the method above was implemented and feature importance ranked by the decrease in mean r2 score computed.

## **Feature Selection Results**

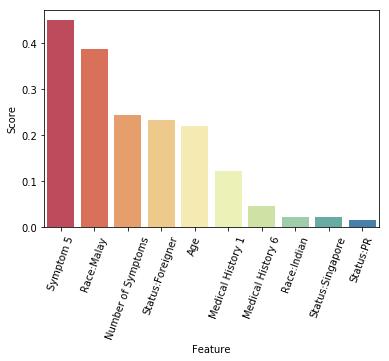


Figure 9. Feature selection bar plot of 10 most important features obtained from feature selection strategy from Extra Trees Regressor model

The 10 most important features were plotted in the figure above for better visualisation. The results show that symptom\_5 (score = 0.449) is the key driver of cost of care, followed closely by Malay racial status (score = 0.388). The next three most important feature are Number of Symptoms, Foreigner Status and Age.

# **Discussion**

### **Statistical Significance and Real-life context of healthcare in Singapore**

The results from the regression model validates the hypothesis from earlier bivariate and correlation analysis, which indicates that the presence of symptom 5 in patients shows the strongest predictive ability for an increase in cost of care. Nonetheless, the results of the feature selection strategy need to be matched with real-life context of healthcare in Singapore.

The relationship between Malay racial status and high cost of care is most likely one of correlation and not causation. In 2014, the Singapore national disease registry revealed that a disproportionate number of patients suffering from either diabetes, myocardiac infarction (MI), stroke and or kidney failure come from the Malay racial group. [3] Some of these diseases, such as stroke and MI may require invasive surgeries which tend to have high costs. These diseases, in turn are associated with higher rates of unhealthy behaviour in the Malay racial group such as smoking. Daily smoking prevalence amongst Malays were the highest and double that of Chinese smokers in 2012.[4] Hence, the underlying driver of cost of care for the Malay racial group could likely be unhealthy behaviour.

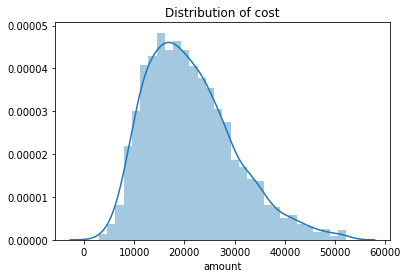
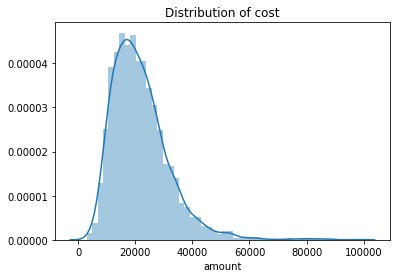
Singapore is a global medical tourism spot and has been awarded the most efficient healthcare system in the world by Bloomberg in 2014. [5] As such, foreigners who come to Singapore to receive healthcare and treatment would be more willing to pay for care. Additionally, foreigners do not receive government subsidies or reimbursements. Hence the lack of subsidies might contribute to the high cost of care that foreigners pay. These reasons might explain the strong correlation between foreigner status and high cost of care. However, these reasonings only hold true on the assumptions that:

1. This dataset is derived from the Singapore healthcare system due to the resident status being classified into Singapore, PR and Foreigner,
2. All 4 amounts for each patient ID and date of admission in bill\_amount.csv can be aggregated as a sum ( instead of a scenario where specific IDs are supposed to be reimbursements and hence deduct from the sum),
3. Subsidies have already been deducted from the amount in bill\_amount.csv

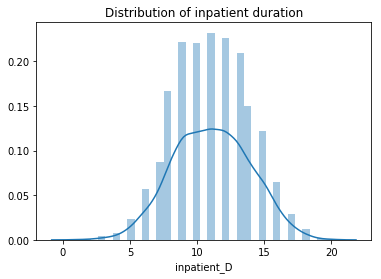
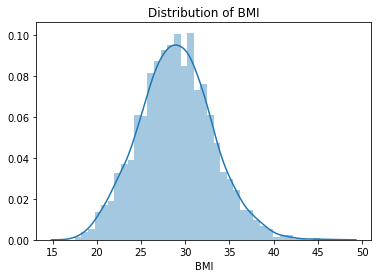
# **Conclusion**

The key driver of cost of care identified from this dataset is the presence of symptom 5. Although Malay racial status is also indicative of higher cost of care, the underlying driver of cost of care most likely stems from unhealthy behaviour prevalent among the racial group. Number of symptoms a patient exhibits is the next strongest driver of cost of care, along with age. Foreigner status indicates higher cost of care, but the underlying context of Singapore as a global medical tourism spot helps to explain this correlation. Notably, the presence of medical history 1 and 6 also seem to be a driver of cost of care.

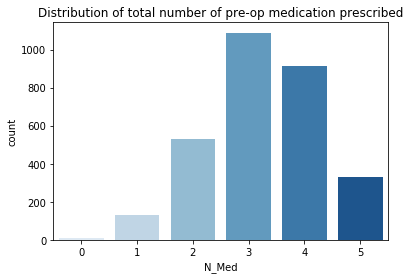
# **Appendix**



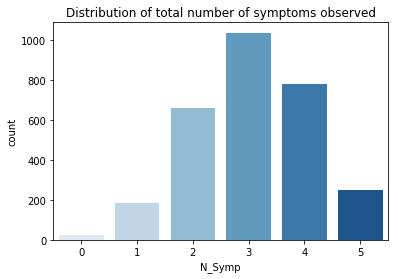
Appendix 1. Distribution of cost before and after outlier removal. Standard deviation reduction



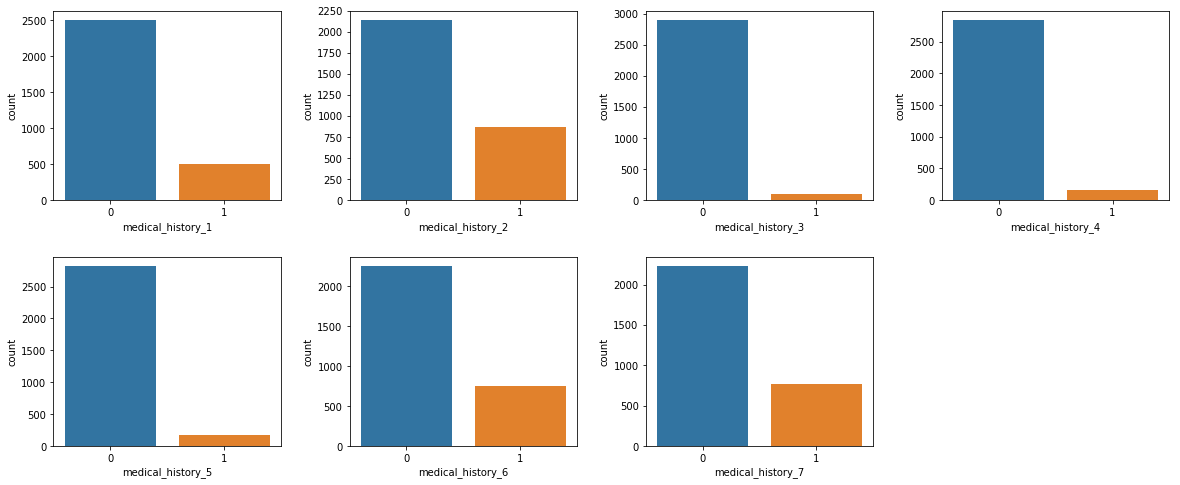
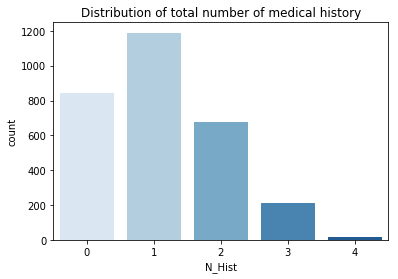
Appendix 2. Distribution plots of BMI and inpatient duration. Both engineered features exhibit a Gaussian distribution.



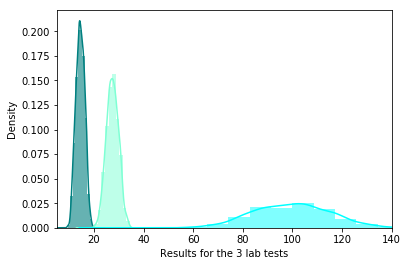
Appendix 3. Distribution of number of preop medication prescribed and count plot of each preop medication. Preop\_medication 1 and 4 were the least frequent medication prescribed amongst the 6, with preop\_medication 5 being present in a majority of samples. This could either be due to higher cost of due to undesired side-effects.



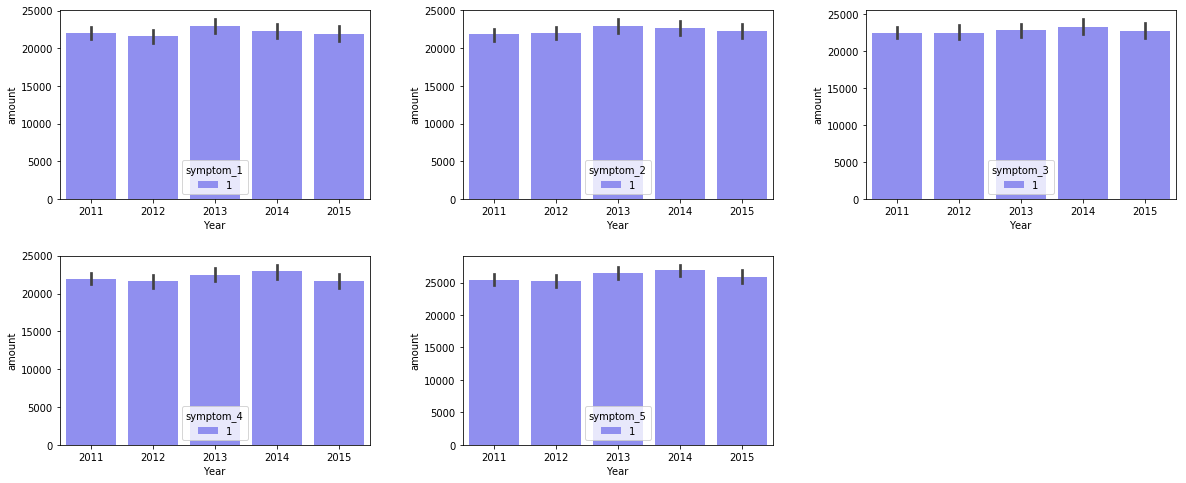
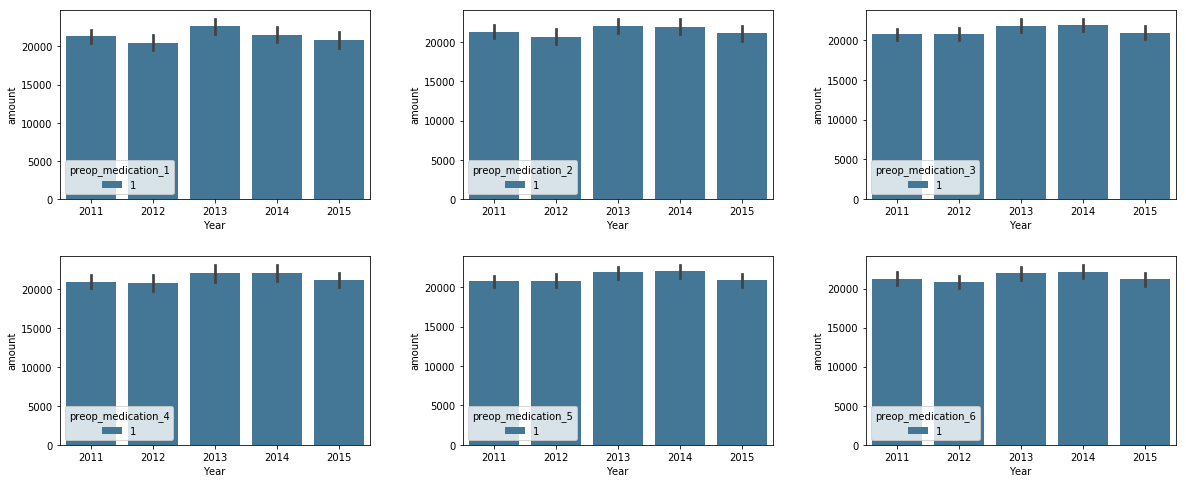
Appendix 4. Distribution of symptoms and count plot for each symptom. Symptom 5 has a lower frequency than the other symptoms.

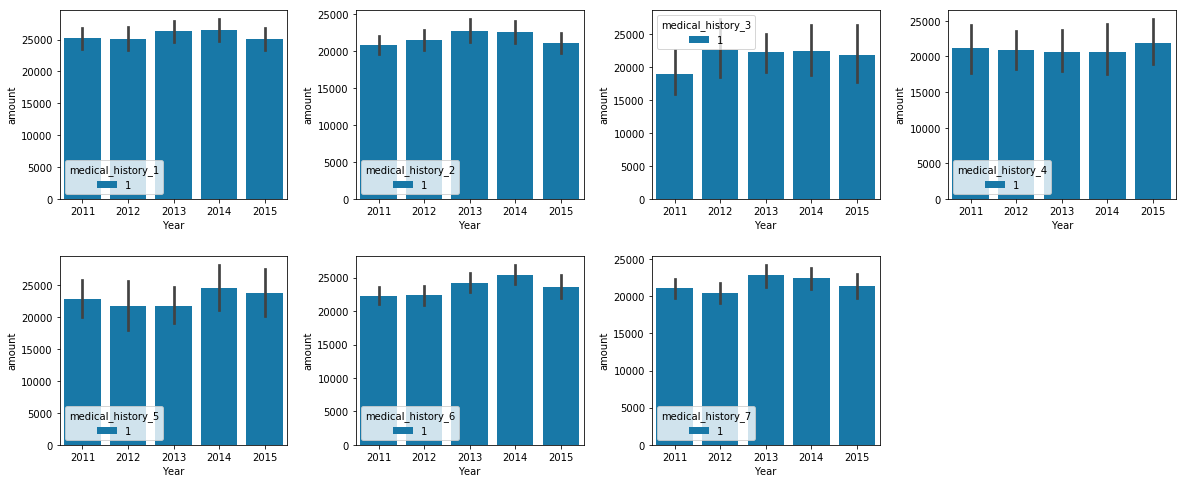


Appendix 5. Distribution of number of medical histroy available and count plot of each medical history. Medical History 2 is the most frequently available among the 7.



Appendix 6. Distribution plot of lab results 1, 2 and 3. All three lab results have different distributions and different means. This indicates that all 3 lab tests are different tests and not replicates. Additionally, lab\_result\_3 suffers from low precision.





Appendix 7. Multivariate analysis of preop medication, medical history and symptom for different years against cost.

# **References**

1. Louppe, G., et al. *Understanding variable importances in forests of randomized trees*. in *Advances in neural information processing systems*. 2013.

2. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**: p. 2825-2830.

3. Khalik, S., *Malay population the most unhealthy group in Singapore*, in *The Straits Times*. 2014: Singapore.

4. Ministry of Health, S., *State of Health - Report of the Director of Medical Services*. 2012.

5. Whitaker, C. *Where Do You Get the Most for Your Health Care Dollar?* Bloomberg Visual Data 2014; Available from: <https://www.bloomberg.com/graphics/infographics/most-efficient-health-care-around-the-world.html>.