Contents

1 Objective

The task is to analyze the clinical and financial data of patients hospitalized for a certain condition. You are required to join the data given in different tables, and find insights about the drivers of cost of care.

2 Data Cleaning

Before analysing, data cleaning has to be done as there were some data which were not coherent. An example would be in the demographics data where the gender, race and resident_status had duplicating categories. In this case, i have edited the names so that the data is coherent. The following are the variables being edited.

- gender: Male, Female
- race: Indian, Malay, Chinese, Others
- resident_status: Singaporean, PR, Foreigner

In the bill data, there were duplicated entries as well. As a result, the total bill amount for a patient had to be combined. In the clinical_data, there were some patients with more than one entry. Hence, i decided to create a column "total_duration" to capture the total hospitalization duration. This can be obtained from the difference between the date of discharge and date of admission. Columns such as "number" was also created to indicate the number of visits the patient made assuming that each bill represents one visit. In total, there are 3000 unique patients in this dataset.

3 Missing Value Imputation

Since there were missing values for *medical_history_2* and *medical_history_5*, imputation had to be done. Categorical variables were converted into dummy variables before performing the imputation using the *mice* package library in R. The method of imputation chosen was through predictive mean matching.

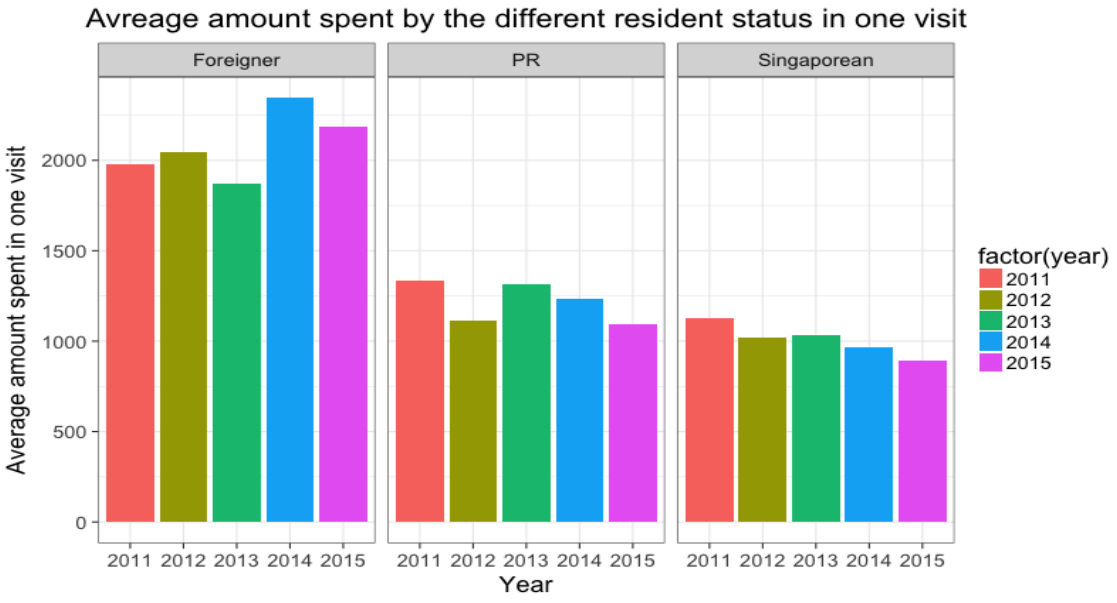
4 Exploratory Data Analysis

4.1 Resident Status V.S Costs

Firstly, i decided to compare the average amount spent by the different resident status in one visit on a yearly basis. I split the data into the respective resident status and computed the total average amount spent per visit.

The total average amount spent per visit is given by the formula:

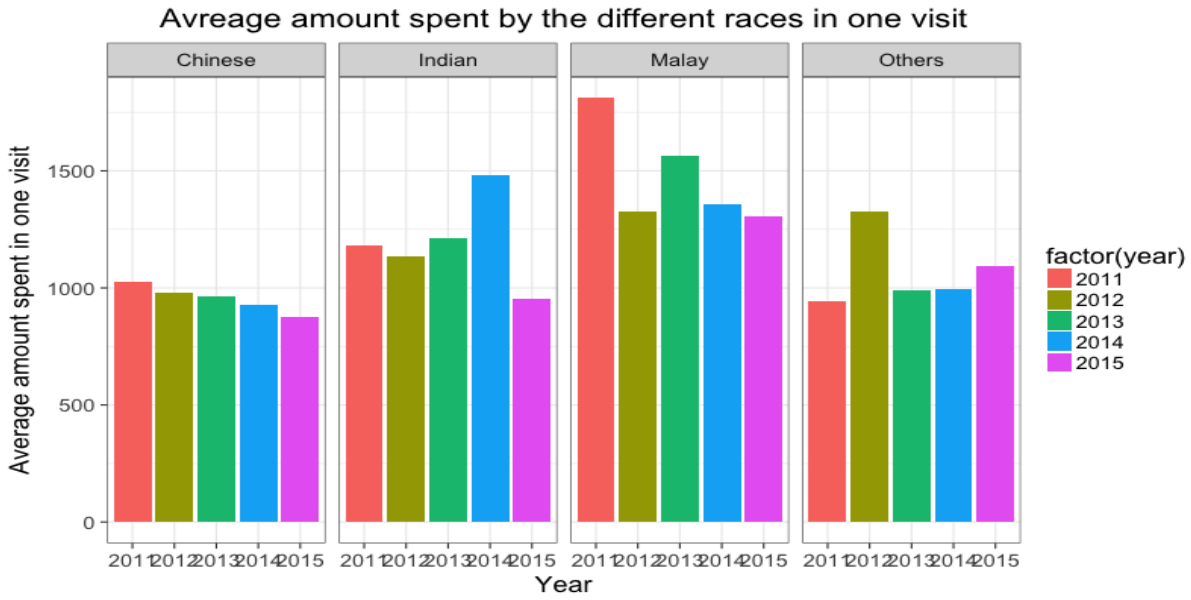
$$\text{Average amount} = \frac{total\_amount}{numberofvisits}$$



We see that foreigners in general incur a much higher costs as compared to PR and Singaporean. There is also a downward trend in general for PR and Singaporeans. On the contrary, the average healthcare costs for foreigners seemed to have increased from 2011 to 2015.

4.2 Ethnicity V.S Costs

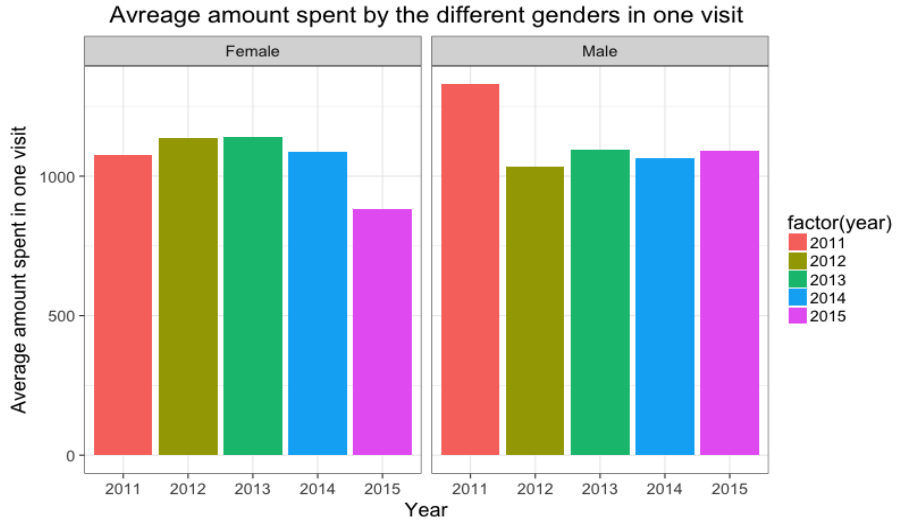
Next, i decided to compare the ethnicity. Similar to resident status, i split the data into the respective ethnicity and computed the total average amount spent per visit.



From the plot, we see that the average healthcare costs has decreased for Chinese, Indian and Malay but increased for Others.

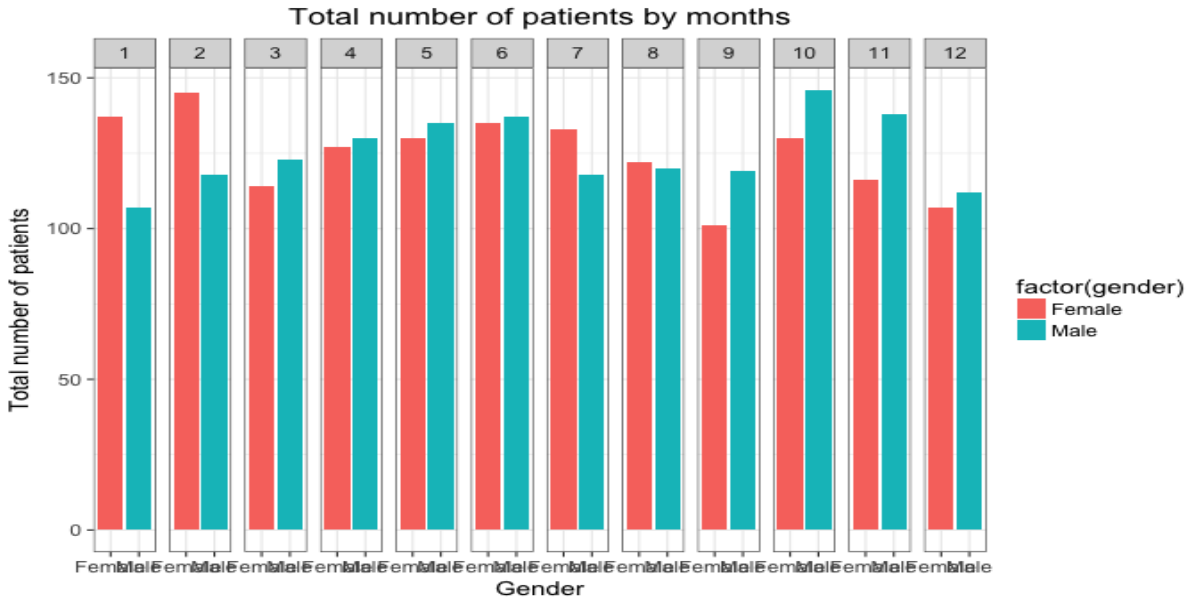
4.3 Gender V.S Costs

Splitting the data into males and females, i computed the total average amount spent per visit for the two different genders. From the plot, we see that healthcare costs is similar between the two genders although it was much higher for males in year 2011 as compared to females.



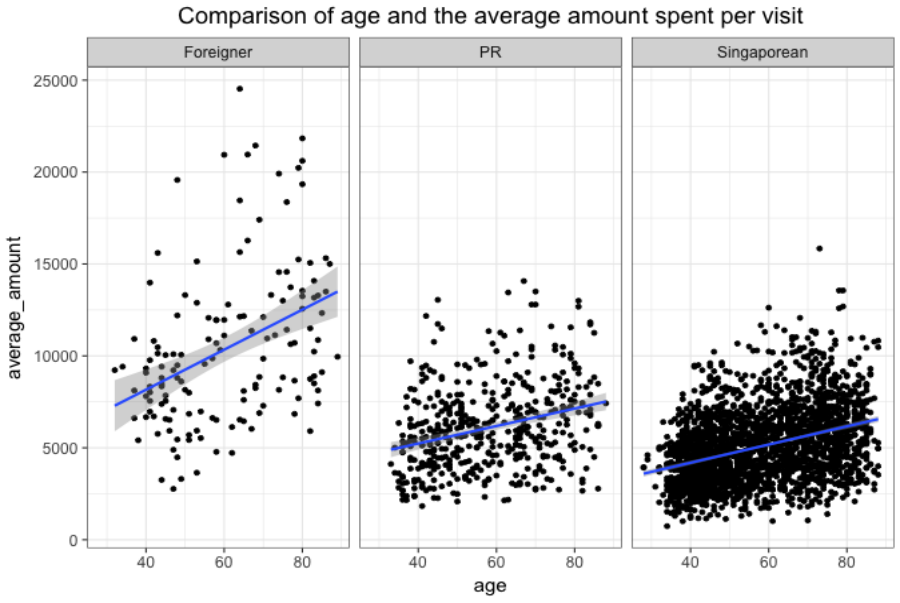
4.4 Identifying peak period

I decided to identify which are the peak months by grouping the data based on their respective months. From the plot, we can see that the lowest number of patient count is during September, December period and the highest number of patient count is during October. Interestingly, there were more female patients in January and February and more male patients in October and November.



4.5 Age V.S Costs

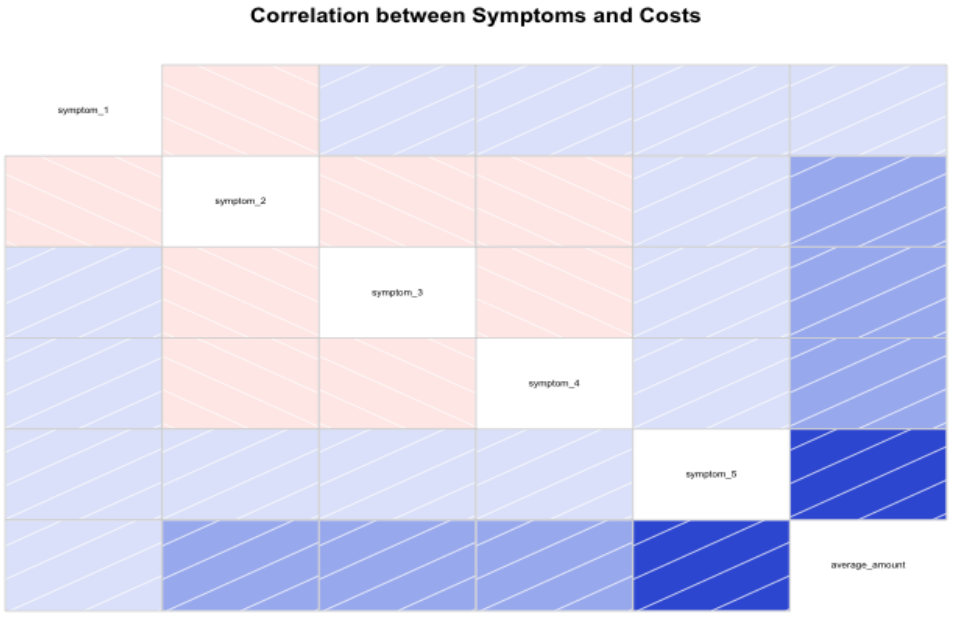
Not only am i interested to see how age can affect the total costs, i am also curious to know how significant it is between the resident status since earlier on we have already identified that foreigners tend to incur higher costs.



From the plot, we see that in general, the average costs increases as age increases. As expected, the costs for foreigners is way higher than PR and Singaporean. It should also be noted that the costs for foreigners are increasing at a much faster rate as represented by the steeper slope.

4.6 Symptoms V.S Costs

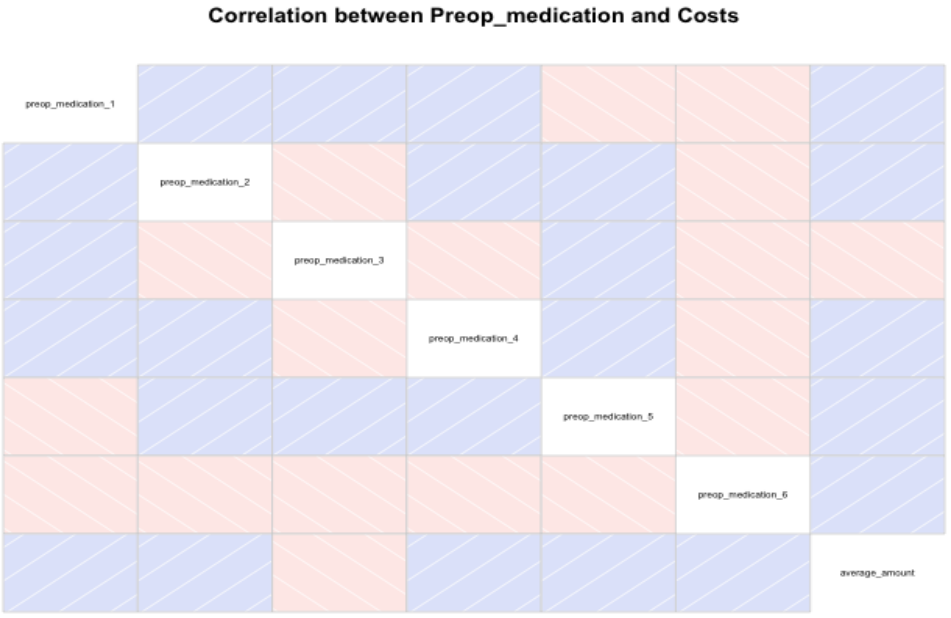
The symptoms of a patient could possibly affect the costs as this might mean more medication required.



From the correlation diagram, all symptoms are positively correlated with the costs. However, it seems that patients with symptom_5 tend to incur much higher costs as compared to other symptoms.

4.7 PreOp-Medication V.S Costs

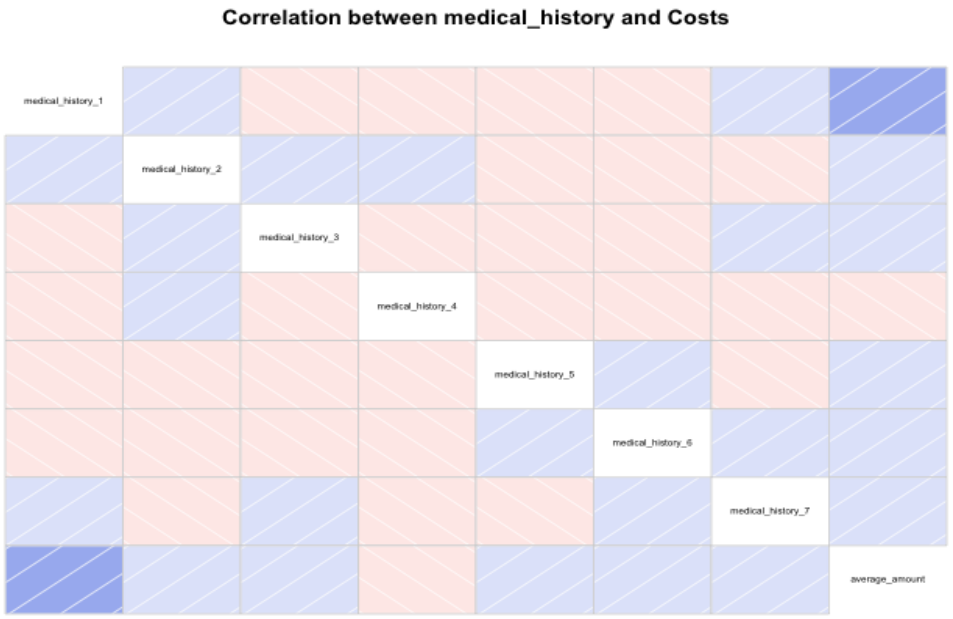
Having pre-operational medication administered could potentially lead to a much higher cost. We can plot a correlation diagram to identify whether they are positively correlated.



From the correlation diagram, it seems that patients with preop_medication1,2,4,5,6 incur higher costs.

4.8 Medical History V.S Costs

Patients with past medical conditions are likely to incur higher costs since they would require more attention. Similarly, we can plot a correlation diagram to identify if they are positively correlated.



From the diagram, patients with medical_history_1 is much more likely to incur higher costs.

4.9 BMI V.S Costs

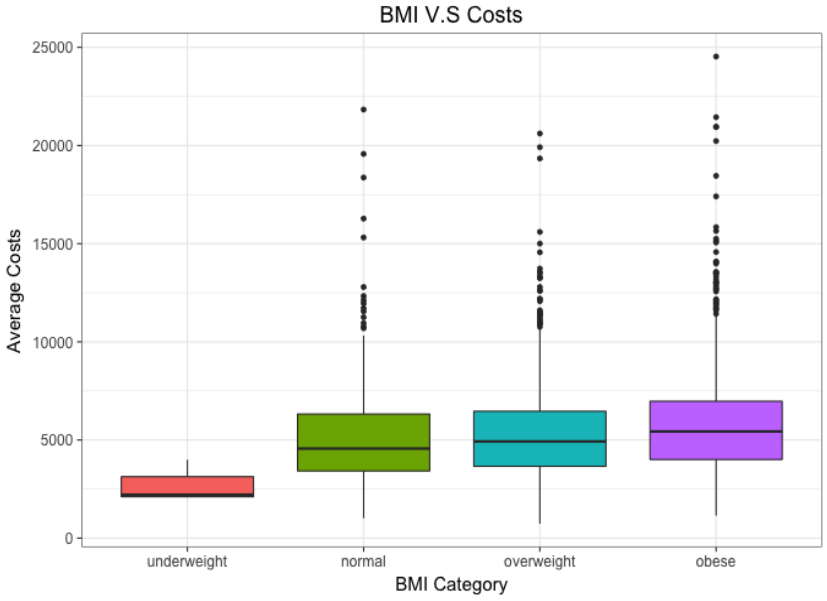
Given the height and weight data of the patients, we can compute the BMI index by the following formula:

$$\text{BMI} = \frac{weight(kg)}{(height(m))^2}$$

In addition, i created another column "BMIcategory" to categorise the patients according to their BMI index.

Category	BMI Index
Underweight	<18.5
Normal	18.5 - 25
Overweight	25 - 30
Obese	>30

Table 1: BMI category



It is not too clear whether there is a significant difference between the costs incurred by obese patients as compared to normal and overweight patients. We can drill in further to conduct a statistical test. First, we have to analyse their respective distribution.

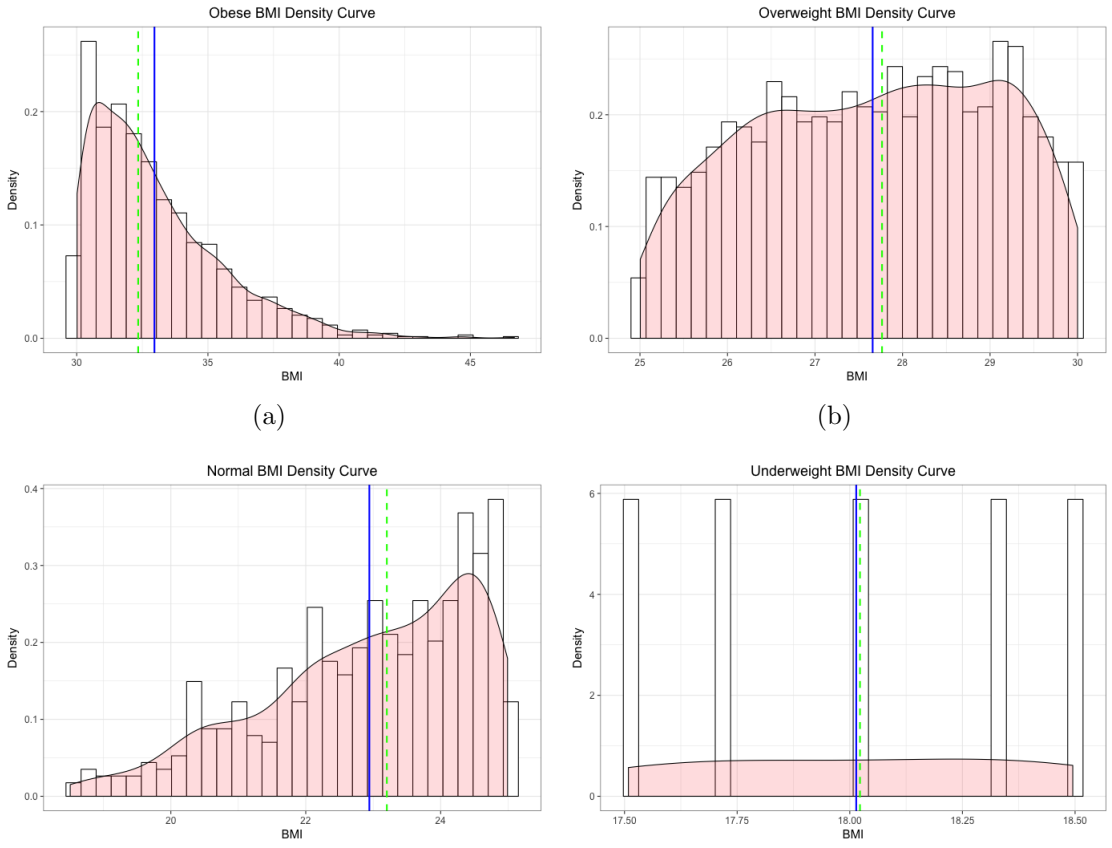


Figure 1: Distributions of various BMIs: (a) Obese; (b) Overweight; (c) Normal; (d) Underweight

Figure 1 shows the respective distributions. The blue solid line represents the mean and the green dashed line represents the median. From the plot, we see that obese, overweight and normal BMIs have skewed distributions.

We are interested to find out if there is a significant difference between the costs for the different BMI groups. Since the distributions are skewed, we will perform Wilcoxon Rank Sum Test, which is a form of non-parametric test. The results are given in Table 2. Taking $\alpha = 0.05$, we con conclude that the costs for different BMIs are significantly different.

	p-value
Obese V.S Overweight	$3.92x10^{-7}$
Obese V.S Normal	$3.39x10^{-9}$
Overweight V.S Normal	0.0243

Table 2: Wilcoxon Rank Sum Test results