Introduction

In the health insurance industry, data analytic is an important process that helps insurers to identify and predict the cost of health insurance which prevents risk and cost for the companies. The Insurance data set contains 1338 observations and 7 variables which include age, sex, children, smoker, region, and expenses.

The description of the data set is as below:
Age: Age of primary beneficiary
Sex: Primary beneficiary's gender
BMI: Body mass index (providing an understanding of the body, weights that are relatively high or low relative to height)
Children: Number of children covered by health insurance / Number of dependents
Smoker: Smoking (yes, no)
Region: Beneficiary's residential area in the US (northeast, southeast, southwest, northwest)
Expenses: Individual medical costs billed by health insurance.

In order for a health insurance company to make money, it needs to collect more in yearly premiums than it spends on medical care to its beneficiaries. As a result, insurers invest a great deal of time and money in developing models that accurately forecast medical expenses for the insured population. The goal of this analysis is to use patient data to estimate the average medical care expenses for such population segments. It is important to give some thought to how these variables may be related to billed medical expenses. For example, we might expect that older people and smokers are at higher risk of large medical expenses. Unlike many other machine learning methods, in regression analysis, the relationships among the features are typically specified by the user rather than being detected automatically. We'll explore some of these potential relationships in the next section.

Methods

Two methods are conducted to analyze the Insurance data set which are multiple linear regression and data mining with regression tree, decision tree in classification to predict the medical expenses based on variables that included.

The information we can determine by using multiple linear regression and data mining are:

- Predict the effect of age, BMI, children, smoker, region to medical expenses.
- How the methods help the insurance company in predicting the healthcare cost?

Multiple Linear Regression

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one dependent variable and two or more independent variables.

Multiple linear regression is selected to research the dataset because according to the dataset's distribution, multiple linear regression is the best choice when there is more than one independent variable used for the prediction of a response variable. For our model, we used smoking, age, BMI, number of children, and region as predictors. Our model's dependent variable is expenses, which measure the medical costs each person charged to the insurance plan for the year.

Data mining: Classification

Data mining basically means digging deep into complicated dataset in different form to find pattern and to translate it to useful information. Large datasets are processed and used with different data mining algorithm to identify patterns and establish relationship between variables to perform data analysis or use the inferred information to make prediction or use it for planning. Classification is one of the famous data mining techniques which involves process of developing a data model that describes and distinguishes data classes and concepts. Classification algorithms does simple job by conducting complicated mathematical computations to identify which subgroup or class does any observation belongs to, based on training dataset. Known category membership from training dataset is mapped with distribution of observed data in training set to

build a decision structure which can further be used as baseline to defining class or category based on data parameters.

We chose to use Classification method among other data mining techniques with dataset we selected because this kind of operation with similar dataset mirrors the real-world use of data mining application. Dataset we are working with is insurance dataset with medical expenses and demographics information of customer of an insurance company. As one of the methods we choose to use regression technique to predict the medical cost of customers, however this technique predicts cost in term of amount. However, for a large company which processes millions of customer data in daily basis it is not quite time efficient to restructure and present among executive leaders for various knowledge discovery. To make decision making process convenient, business would like to classify their customer in different classes based on their medical expenses. Classification techniques maps each expense classes with range of demographics, this classification of observation of customer information into level of expense certain customers are likely to incur can be useful for company to make different business decision.

Using this technique, business like insurance company will be able to classify the population based on what probable medical cost, this information help business as such to make different decision as such.

1. Which expense class demographics should be targeted to generate maximum revenue?
2. Decide premium rate based on predicted expenses class
3. Discover which community or geographical region is less risk.
4. Calculate cost factor which need to be applied on insurance premium of a customer who is smoker.

Analysis and Interpretation

1. Exploration of the data set and visualization of the data:

As part of the preliminary analysis, we will conduct exploratory data analysis on Insurance dataset.

We can start exploring data by observing the structure of dataset.

```
> str(insu)
'data.frame':                1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ expenses: num  16885 1726 4449 21984 3867 …
```

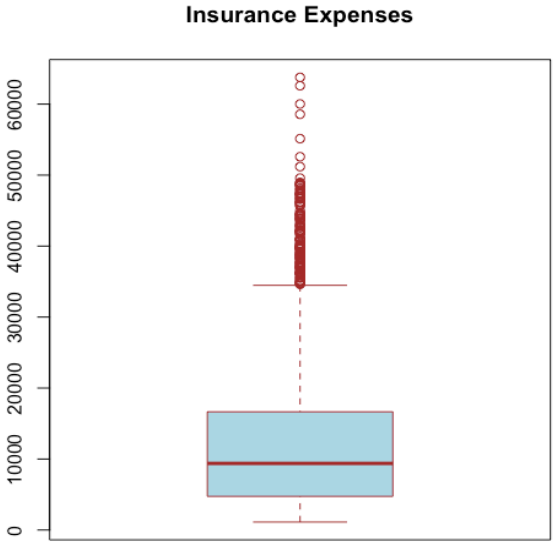
We have 1338 record of customer with 7 variables. Variable “sex” and “smoker” are factors with 2 levels and “region” is a factor with 4 levels. Variable “expenses” will be considered as dependent variable which is influenced by the rest of the variable in data which are considered as independent variables.

We will explore the distribution of each parameter using summary () function.

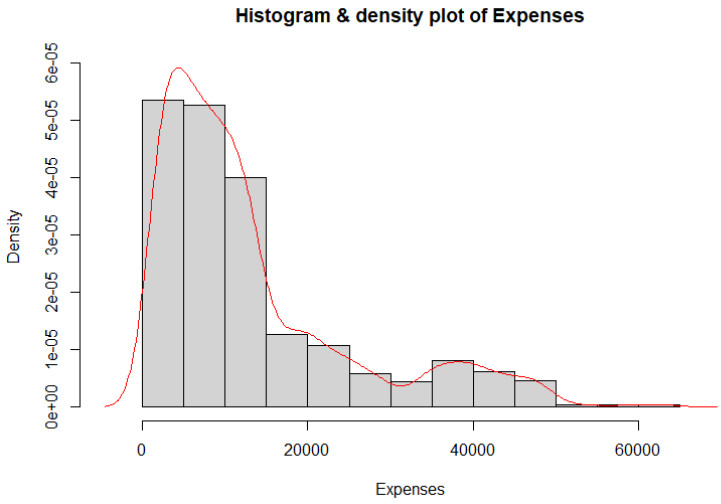
```
> summary(insu)
      age            sex                bmi          children       smoker            region          expenses
 Min.   :18.00   Length:1338        Min.   :16.00   Min.   :0.000   Length:1338        Length:1338        Min.   : 1122
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character   1st Qu.: 4740
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character   Median : 9382
 Mean   :39.21                      Mean   :30.67   Mean   :1.095                                         Mean   :13270
 3rd Qu.:51.00                      3rd Qu.:34.70   3rd Qu.:2.000                                         3rd Qu.:16640
 Max.   :64.00                      Max.   :53.10   Max.   :5.000                                         Max.   :63770
```

Sample data has age distribution from 18 to 64-year-old customer, this is a good range of age group to train model. Focusing on expenses, we can see max expense is screw away from mean, median and 3rd quartile, this suggests there can be outliers. We will visualize the interquartile distribution of expenses using boxplot and histogram.

```
> IQR(insu$expenses)
[1] 11899.63
> range(insu$expenses)
[1]  1121.87 63770.43
> boxplot(insu$expenses)
```



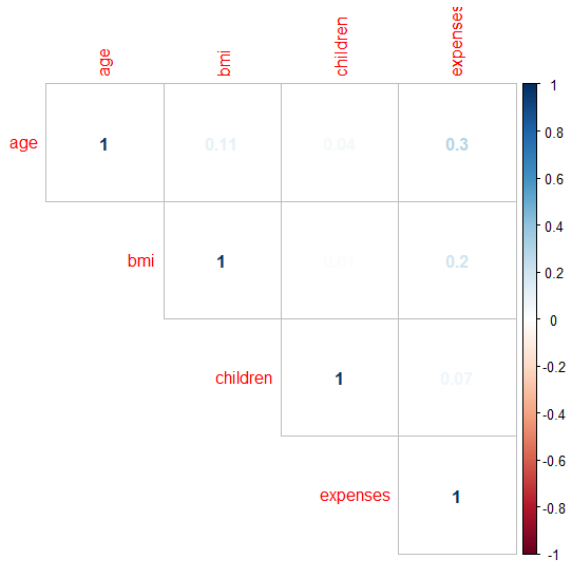
The above boxplot suggests the frequency distribution of expenses is not distributed normally but more skewed towards small value of expenses. To confirm, we will plot a histogram.



Since the frequency distribution of expenses is not equally distributed, we might observe bias in our model trained by this dataset. To obtain better accuracy of the model, we will have to remove outliers.

Now, we will explore the correlation between each parameter, two independent parameters strongly correlated to each other may cause duplication in the model. First, we will plot the correlation for the numerical parameter of the dataset.

```
> cor <- cor(select(ins,age,bmi,children, expenses))
> corrplot(c,method = "number", type = "upper")
```



In the above diagram, we can observe a positive correlation between expense and age and very small with children as well, but no strong correlation between independent variable as expected.

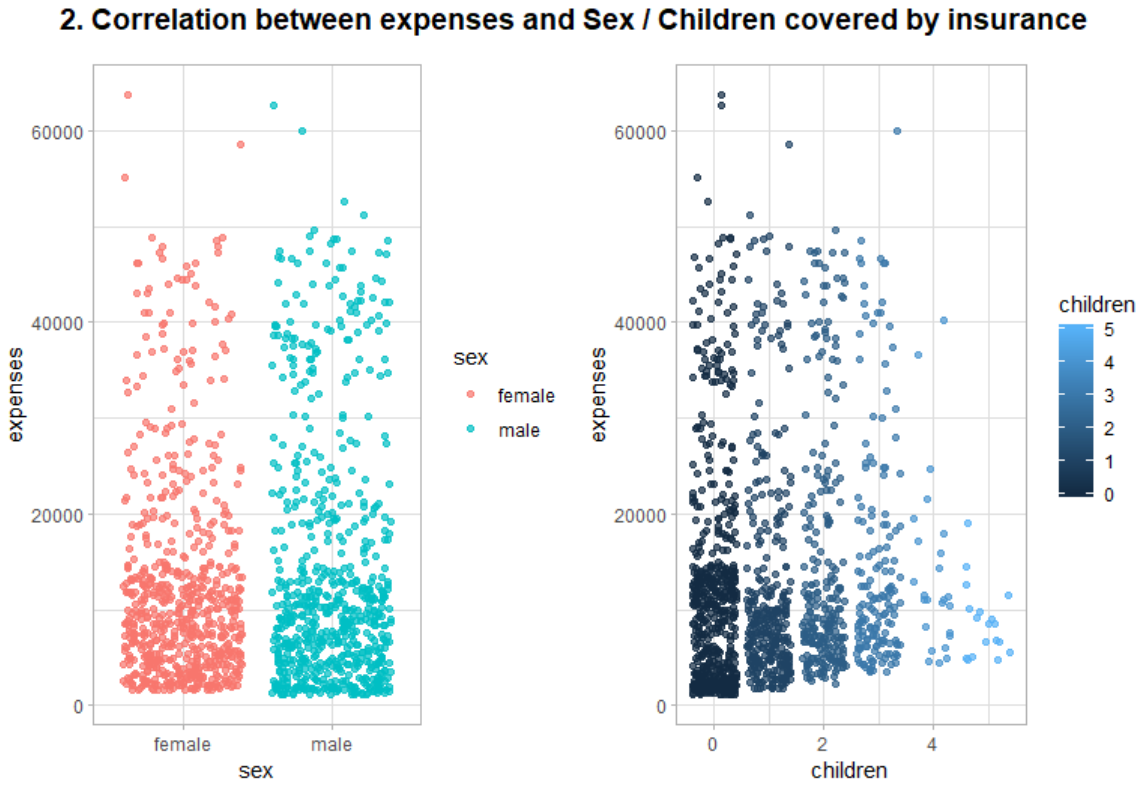
To visualize correlation between expense and categorical parameter, we will use ggplot.

```
> # Correlation between Expenses and Age / BMI
> x <- ggplot(insu, aes(age, expenses)) +
+   geom_jitter(color = "blue", alpha = 0.5) +
+   theme_light()
> y <- ggplot(insu, aes(bmi, expenses)) +
+   geom_jitter(color = "green", alpha = 0.5) +
+   theme_light()
> p <- plot_grid(x, y)
> title <- ggdraw() + draw_label("1. Correlation between Expenses and Age / BMI",
+                                fontface='bold')
> plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))
```

As Age and BMI go up Expenses for health insurance also trends up. Very strong linear relation can be observed between expenses and age, from the above graph we conclude expense increase with an increase in age, also similar not as strong pattern can be observed in graph between expenses and BMI. It can be said, BMI has a weak positive correlation with expenses.

```
> # Correlation between Expenses and Sex / Children covered by insurance
> x <- ggplot(ins, aes(sex, expenses)) +
+   geom_jitter(aes(color = sex), alpha = 0.7) +
+   theme_light()
> y <- ggplot(ins, aes(children, expenses)) +
+   geom_jitter(aes(color = children), alpha = 0.7) +
+   theme_light()
> p <- plot_grid(x, y)
> title <- ggdraw() + draw_label("2. Correlation between expenses and Sex / Children covered by insurance",
+                                fontface='bold')
```

```
> plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))
```



In the above plot of expense vs children, no pattern can be observed in the above diagram.

Also, with respect to expenses against sex, we can observe that expenses for females and males both are distributed, however, we cannot answer if the gender of a patient or customer influences the medical expense. We can conduct hypothesis testing or regress sex with expenses to identify the influence of gender on expenses.

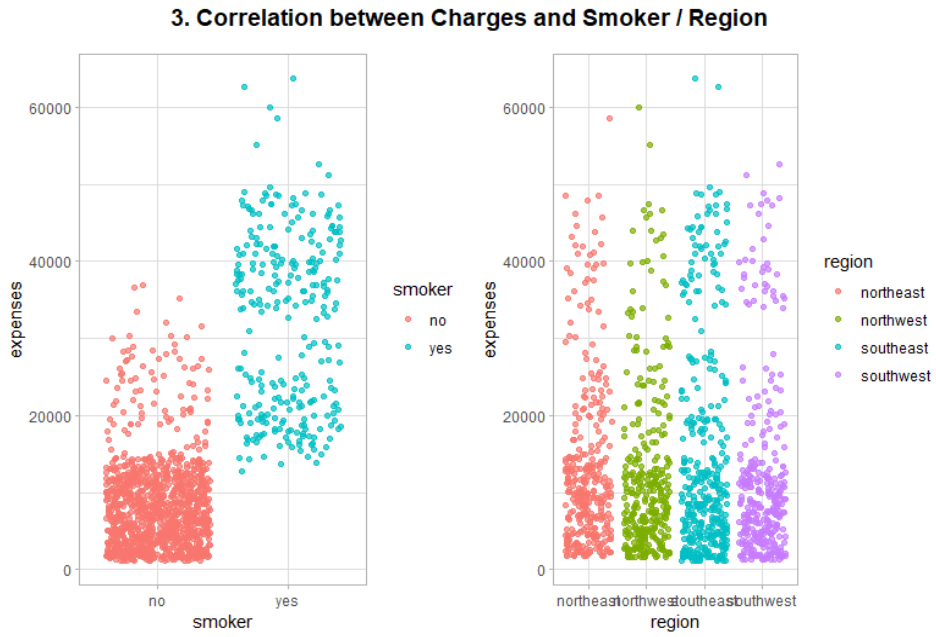
We will regress sex against expenses to identify the statistical significance of gender in the medical expenses of customer based on this dataset.

```
> summary(model.sex)$coeff
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 12569.579   470.0717 26.739706 1.626105e-126
sexmale      1387.173   661.3309  2.097547  3.613270e-02
```

The coefficient from the above model suggests that sex of a customer significantly influences the medical expenses. Based on our sample dataset, the average expenses of a male is 1387 more than that of female. Also, since p-value of dummy variable “sexmale” is significantly below level 0.05, this suggests a strong relation between sex and expenses.

Now we will plot to visualize correlation between Expense and smoke/region

```
> #Correlation between Expense smoke/region
> x <- ggplot(insu, aes(smoker, expenses)) +
+   geom_jitter(aes(color = smoker), alpha = 0.7) +
+   theme_light()
> y <- ggplot(insu, aes(region, expenses)) +
+   geom_jitter(aes(color = region), alpha = 0.7) +
+   theme_light()
> p <- plot_grid(x, y)
> title <- ggdraw() + draw_label("3. Correlation between Charges and Smoker / Region", fontface='bold')
> plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))
```



From above diagram, it is clearly visible that the medical expenses of a customer are highly influenced by the fact they smoke or not.

However, in expense vs region plot, we cannot visually infer if the region of customer makes any difference in their medical cost. To identify the statistical significance of geographical region of customer in their medical expenses, we will regress expenses vs region.

```
> model.region<- lm(expenses ~ region, data = insu)
> summary(model.region)$coeff
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept)     13406.3847   671.2966 19.970880 7.671566e-78
regionnorthwest  -988.8095   948.6263 -1.042359 2.974339e-01
regionsoutheast  1329.0268   922.9071  1.440044 1.500896e-01
regionsouthwest -1059.4468   948.6263 -1.116822 2.642715e-01
```

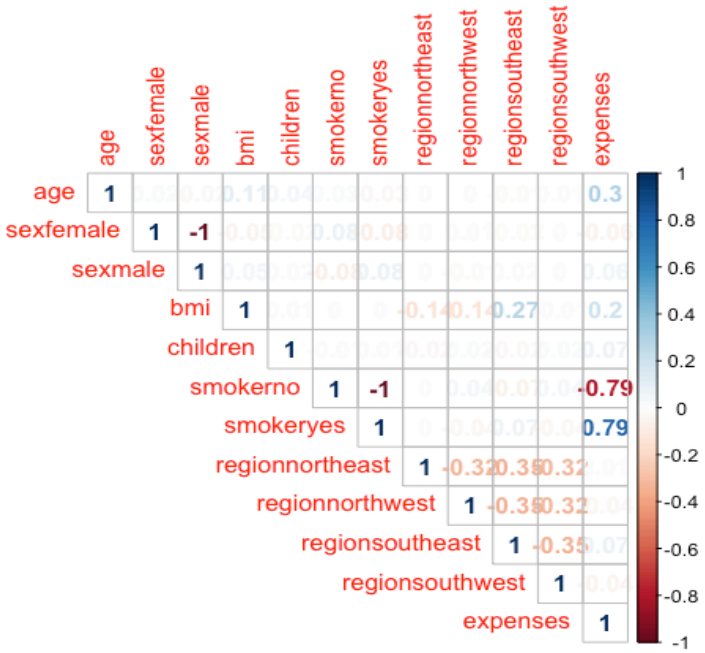
Coefficient of expenses vs region model also suggests that region of customer is statistically significant to influence the medical expenses, all of the regions p-value are below 0.05.

From above graphical and mathematical exploration, we can conclude that all the parameter is influential to medical expenses of a customer, so we will be including each of them to train models. We noticed, smokers as a strong factor followed by age, BMI, sex, and region. Children are found to be a weak positive correlation.

Now for the compatibility of different modeling techniques and functions, we will change all the factor parameters such as sex, smoke, and region to numerical indicator.

To do so, we will use dummyVars() function from “caret” package.

```
install.packages("caret")
library(caret)
dmy <- dummyVars("~.", data = insu)
insurance <- data.frame  (predict(dmy, newdata=insu))
```



As we can see from the correlation plot above, the correlation between the smoker and expenses is strong, it shows that expenses for Smokers are higher for non-smokers.

2. Multiple linear regression

Next step, a multiple linear regression model is conducted to predict the insurance cost.

First, we prepare and split the data: We split our dataset into a training set and a testing set. The training set is used to train model and the test set is used to evaluate the model.

```
n_train <- round(0.8 * nrow(insu))
train <- sample(1:nrow(insu), n_train)
x_train <- insu[train, ]
x_test <- insu[-train, ]
```

Train and test model: After we have a train and test datasets, we can evaluate the model using Linear regression as below:

```
> model_0 <- lm(formula = expenses~., data = x_train)
> summary(model_0)

Call:
lm(formula = expenses ~ ., data = x_train)

Residuals:
     Min       1Q   Median       3Q      Max
-11421.6  -2990.9   -972.4   1444.1  30108.3

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11943.52    1102.50 -10.833   <2e-16 ***
age                263.19      13.38  19.667   <2e-16 ***
sexmale           -202.13     377.43  -0.536   0.5924
bmi                337.73      32.48  10.399   <2e-16 ***
children           384.02     157.86   2.433   0.0152 *
smokeryes        23668.38     462.60  51.164   <2e-16 ***
regionnorthwest   -572.37     541.69  -1.057   0.2909
regionsoutheast  -1148.78     537.53  -2.137   0.0328 *
regionsouthwest   -799.97     533.18  -1.500   0.1338
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6131 on 1061 degrees of freedom
Multiple R-squared:  0.7504,    Adjusted R-squared:  0.7485
F-statistic: 398.7 on 8 and 1061 DF,  p-value: < 2.2e-16
```

As we can see the variable “sex” have less significance values while smoking seems to have a huge influence on expenses. The results of the linear regression model make logical sense: old age, smoking, and obesity tend to be linked to additional health issues, while additional family member dependents may result in an increase in physician visits and preventive care such as vaccinations and yearly physical exams. However, we currently don’t know how well the model is fitting the data. Here the R-squared value for this model is 0.7504. We will create one more model eliminating the column sex and check if performance can be improved.

Train and Test New Model:

```
> model_1 <- lm(formula = expenses ~. - sex, data = x_train)
> summary(model_1)

Call:
lm(formula = expenses ~ . - sex, data = x_train)

Residuals:
     Min       1Q   Median       3Q      Max
-11518.5  -2968.3   -976.8   1414.9  30026.7

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -12009.53    1095.23 -10.965   <2e-16 ***
age                263.26      13.38  19.679   <2e-16 ***
bmi                336.67      32.41  10.389   <2e-16 ***
children           383.94     157.81   2.433   0.0151 *
smokeryes        23645.02     460.38  51.360   <2e-16 ***
regionnorthwest   -572.01     541.50  -1.056   0.2911
regionsoutheast  -1150.60     537.34  -2.141   0.0325 *
regionsouthwest   -797.70     532.98  -1.497   0.1348
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6129 on 1062 degrees of freedom
Multiple R-squared:  0.7503,    Adjusted R-squared:  0.7487
F-statistic:   456 on 7 and 1062 DF,  p-value: < 2.2e-16
```

Comparing the models:

```
> print(paste0("R-squared for first model:", round(r_sq_0, 4)))
[1] "R-squared for first model:0.7504"
> print(paste0("R-squared for new model: ", round(r_sq_1, 4)))
[1] "R-squared for new model: 0.7503"
> print(paste0("RMSE for first model: ", round(rmse_0, 2)))
[1] "RMSE for first model: 5799.94"
> print(paste0("RMSE for new model: ", round(rmse_1, 2)))
[1] "RMSE for new model: 5797.75"
```

The multiple R-squared value which is known as the coefficient of determination provides a measure of how well our model as a whole explains the values of the dependent variable. It is

similar to the correlation coefficient, in that the closer the value is to 1.0, the better the model perfectly explains the data. Since the R-squared value is 0.7503, we know that the model explains more than 75 percent of the variation in the dependent variable. It is useful for comparing the performance of models with different numbers of explanatory variables. From the result above, the performance is quite similar between two models so we will keep the new model since it's a little bit simpler.

Model Perform Prediction:

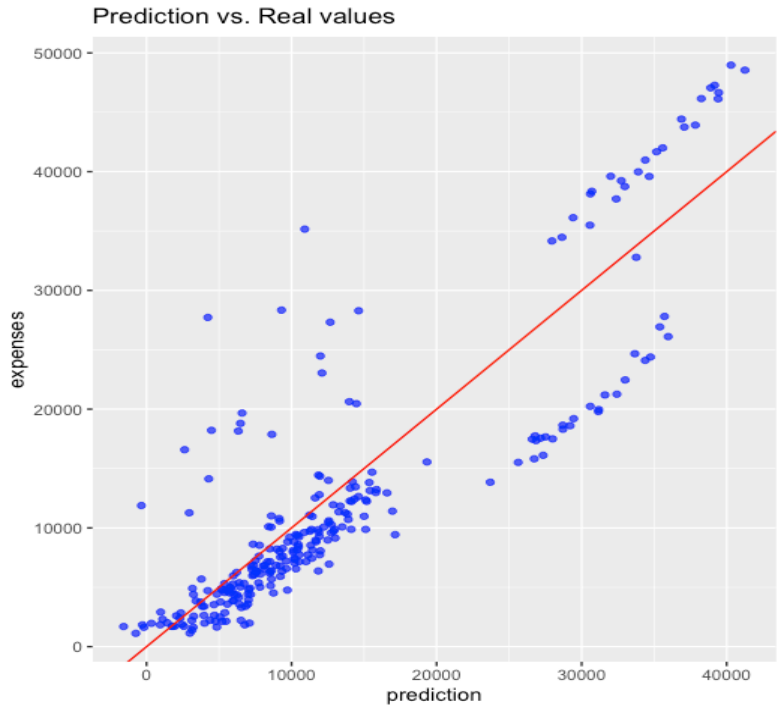
```
> predicted <- predict(object=model_0,newdata=x_test, type = "response")
> result.df <- data.frame(cbind(actuals = x_test$expenses, predicted=predicted))
> result.df <- result.df %>%
+   mutate(error = result.df$actuals - result.df$predicted )%>%
+   round(.,2)
> result.df <- result.df %>%
+   mutate(error_percent = paste0(round(result.df$error/result.df$actuals*100,2),"%"))
> kable(head(result.df))
```

```
|  actuals| predicted|    error|error_percent |
|--------:|---------:|--------:|:-------------|
| 16884.92|  26159.67| -9274.75|-54.93%       |
|  4449.46|   6071.25| -1621.79|-36.45%       |
|  7281.51|   7895.97|  -614.46|-8.44%        |
| 27808.73|  36152.71| -8343.98|-30%          |
|  1837.24|    719.51|  1117.73|60.84%        |
| 13770.10|  13501.37|   268.73|1.95%         |
> sprintf("The Average percent error is: %s%%",
+         round(mean(result.df$error/result.df$actuals*100),2))
[1] "The Average percent error is: -8.73%"
```

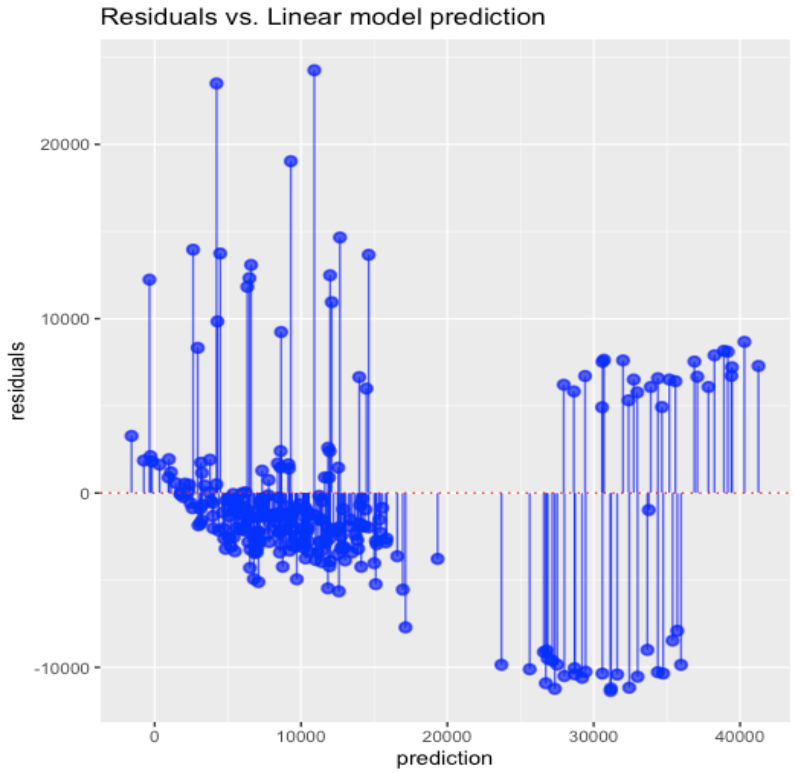
The above results show us the error difference between the actual and predicted values and the error percentage. Our model was able to predict the premium insurance for policyholders with a mean difference of ~8.73%.

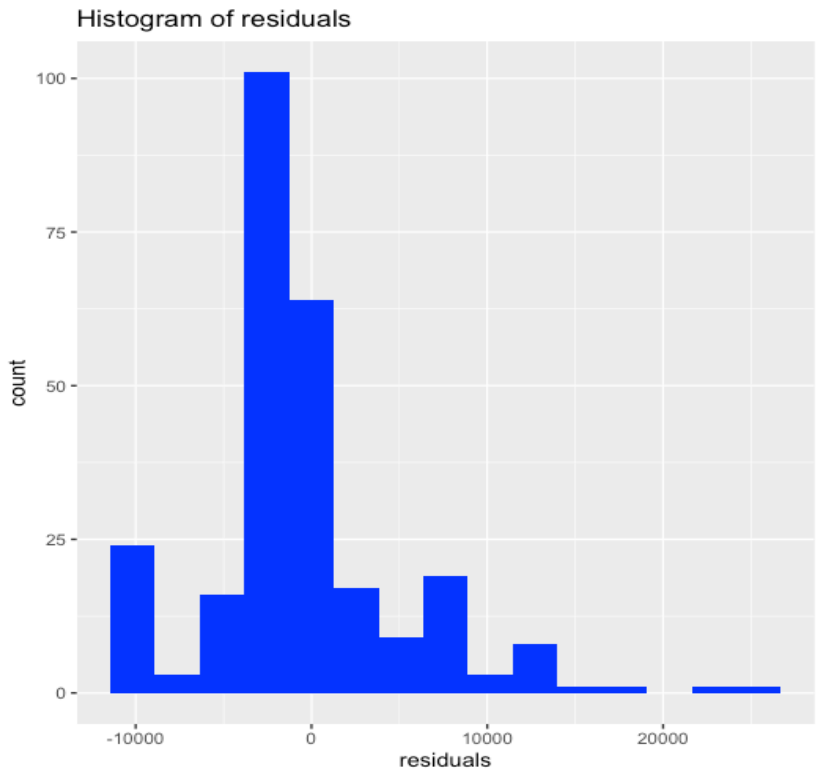
We compare the prediction results with actual results by plotting as below:

```
> x_test$prediction <- predict(model_1, newdata = x_test)
> ggplot(x_test, aes(x = prediction, y = expenses)) +
+   geom_point(color = "blue", alpha = 0.7) +
+   geom_abline(color = "red") +
+   ggtitle("Prediction vs. Real values")
```




```
> x_test$residuals <- x_test$expenses - x_test$prediction
> ggplot(data = x_test, aes(x = prediction, y = residuals)) +
+   geom_pointrange(aes(ymin = 0, ymax = residuals), color = "blue", alpha = 0.7) +
+   geom_hline(yintercept = 0, linetype = 3, color = "red") +
+   ggtitle("Residuals vs. Linear model prediction")
> ggplot(x_test, aes(x = residuals)) +
+   geom_histogram(bins = 15, fill = "blue") +
+   ggtitle("Histogram of residuals")
```


Residuals vs. Linear model prediction
residuals
20000
10000
0
-10000
0
10000
20000
30000
40000
prediction


Histogram of residuals
count
100
75
50
25
0
-10000
0
10000
20000
residuals

We can see the errors in the model are close to zero so model predicts well. In conclusion, we can conclude from the above results that the variable “Smoker” is highly correlated with “Expenses” which means a Smoker is very likely to hold premium insurance.

Applying the model on new data:

For example, we apply and test 2 different people and see what expenses on health care will be for them.

Joe: 34 years old, BMI 26.5, male, has no children, smokes, from northwest region.

Kay: 30 years old, BMI 31.2, female, 2 children, doesn’t smoke, from southeast region.

```
> Joe <- data.frame(age = 34,
+                   sex= "male",
+                   bmi = 26.5,
+                   children = 0,
+                   smoker = "yes",
+                   region = "northwest")
> print(paste0("Health care charges for Joe: ", round(predict(model_1, Joe), 2)))
[1] "Health care charges for Joe: 29836.39"
> Kay <- data.frame(age = 30,
+                   sex = "female",
+                    bmi = 31.2,
+                    children = 0,
+                    smoker = "no",
+                    region = "northeast")
> print(paste0("Health care charges for Kay: ", round(predict(model_1, Kay), 2)))
[1] "Health care charges for Kay: 6559"
```

From the result of this example above, the use of multiple linear regression model help insurance company predicts the health insurance for their patient. Patient data such as age, sex, BMI, smoke or no smoke, region estimates the average medical care expenses for such population segments. It shows how these variables related to insurance expenses.

3. Data Mining using Decision tree (Classification)

We will use above prepared "insurance" data frame for regression. However, for datamining by classification algorithms such as decision tree, we will have to transform our outcome variable "expenses" to levels or categories.

To do so, we will replicate "insurance" data frame to "imining" data frame.

```
imining <- insurance
```

Now, based on range and distribution of expenses, we will define category.

Expenses Level	Expenses ($)
L1	Below 3000
L2	3000 to 6000
L3	6000 to 9000
L4	9000 to 12000
L5	12000 to 15000
L6	15000 to 18000
L7	180000 to 36000
L8	36000 to 72000
L9	Above 72000

We will change number value of expenses to categories as specified in above table.

```
imining$expenses <- cut(imining$expenses,
+      breaks = c(0, 3000, 6000, 9000, 12000, 15000, 18000, 36000,72000, 100000),
+     labels = c("L1","L2","L3", "L4", "L5", "L6", "L7", "L8", "L9"), right = F)
```

Let us look new structure of "imining" dataset now.

```
> str(imining)
'data.frame':                1338 obs. of  12 variables:
 $ age           : num  19 18 28 33 32 31 46 37 37 60 ...
 $ sexfemale     : num  1 0 0 0 0 1 1 1 0 1 ...
 $ sexmale       : num  0 1 1 1 1 0 0 0 1 0 ...
 $ bmi           : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
 $ children      : num  0 1 3 0 0 0 1 3 2 0 ...
 $ smokerno      : num  0 1 1 1 1 1 1 1 1 1 ...
 $ smokeryes     : num  1 0 0 0 0 0 0 0 0 0 ...
```

```
 $ regionnortheast: num  0 0 0 0 0 0 0 0 1 0 ...
 $ regionnorthwest: num  0 0 0 1 1 0 0 1 0 1 ...
 $ regionsoutheast: num  0 1 1 0 0 1 1 0 0 0 ...
 $ regionsouthwest: num  1 0 0 0 0 0 0 0 0 0 ...
 $ expenses       : Factor w/ 9 levels "L1","L2","L3",..: 6 1 2 7 2 2 3 3 3 7 ...
```

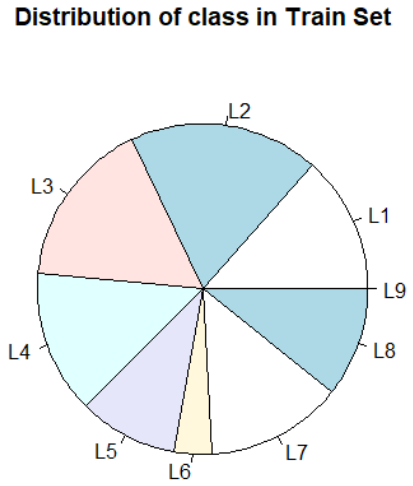
Now, we have our dataset ready for classification data mining with dependent variable as factor. We will install and load the library required for classification.

```
library(party)
install.packages("caret")
library(caret)
```

We will divide dataset into training set and testing set, we will set seed before it to make sure the result can be reproduced if desired. If will divide dataset into train and test by 7:3 proportion.

```
set.seed(123)
ind <- sample(2, nrow(imining), replace=TRUE, prob = c(0.70,0.30))
train.data <- imining[ind==1,]
test.data <- imining[ind==2,]
```

To make sure the model is trained well, the sample of each class should be proportionally distributed or at least there should not be very high variance in distribution of data. To visualize the proportionality, we will plot a pie chart.



We can see that train set has good proportion of each class except “L6” and “L9”, this is because we have defined “L9” class as above max of our training set, if the data parameter does not meet parameter of every other class, it will be assigned class “L9” which is extremely high expenses.

And, the reason we have few samples for L6 class in train set is because of actual class distribution of L6 is less in our original dataset.
The below table is class distribution of original dataset.

L1	L2	L3	L4	L5	L6	L7	L8	L9
191	238	219	198	134	45	186	127	0

To train our model, we will now define formula for class variable of the model “expenses”. Since, all the parameters of customers are already changed into numerical value, we can include every observed implement of dataset in our formula. All the element in dataset can be denotated directly by a period “.”

Formula:

```
expen.form <- expenses ~ .
```

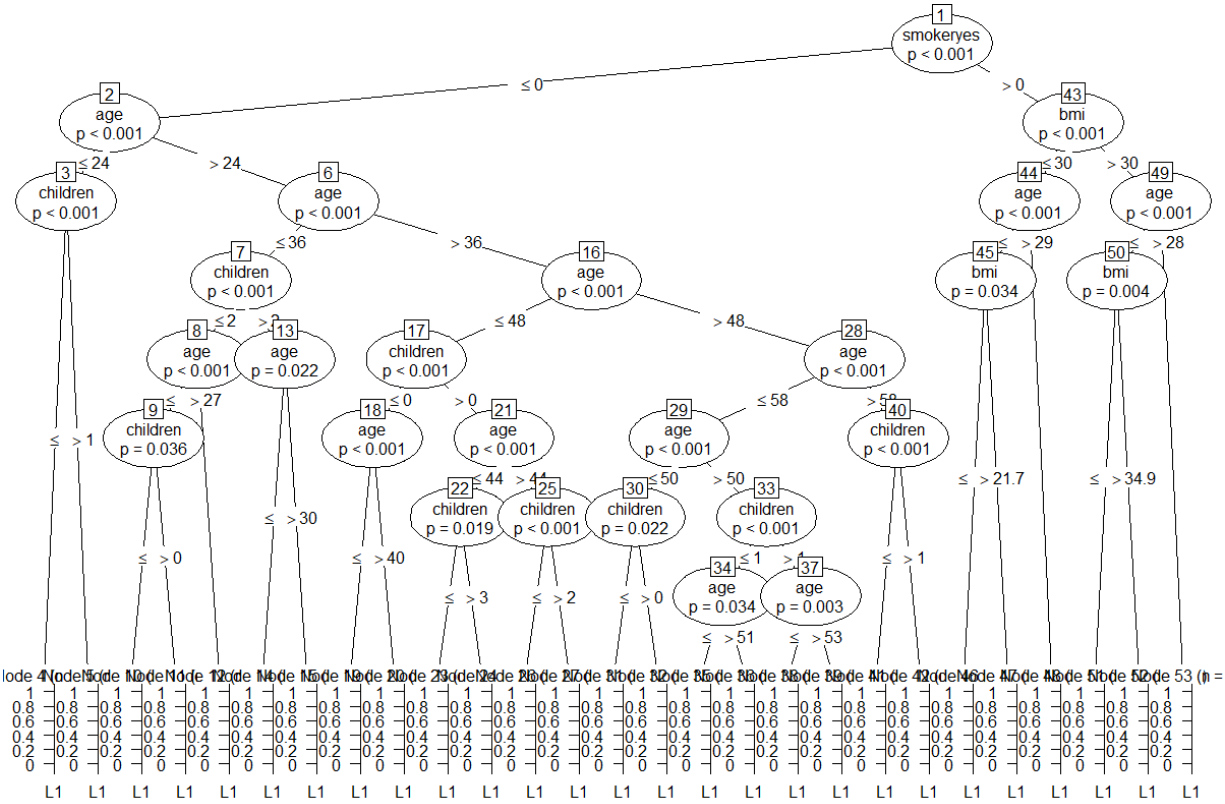
The above formula is a symbolic description of the model to be fit and is equivalent to
Expense = age + sexmale + smokeryes + BMI + children + regionnortheast + regionnorthwest + regionsoutheast + regionsouthwest

For binary factor parameters such as smoker and sex, only one of the dummy variables can provide information if the record is of male or female and smoker and non-smoker.
Now, we can generate conditional inference trees estimate of class based on train set sample and formula. We will plot the decision tree to visualize.

```
expen_ctree <- ctree(expen.form, data = train.data)
plot(expen_ctree)
```

Although, we are not able to view all the nodes clearly in the above diagram, but we can clearly see that smoker is the most important factor followed by age and BMI. Now that we have our

decision tree model ready, we can use it to predict the class of medical expense based on customer demographics.



First, we will test this model with data set which we used to train the model. We visualize the accuracy of the model; we will generate confusion matrix.

```
train.predict <- predict(expen_ctree)
confusionMatrix(train.predict, train.data$expenses)
```

```
> confusionMatrix(train.predict, train.data$expenses)
Confusion Matrix and Statistics

          Reference
Prediction  L1  L2  L3  L4  L5  L6  L7  L8  L9
        L1 121   9   1   3   3   2   4   0   0
        L2   6 155   1   1   2   3  15   0   0
        L3   0  12 149   7   3   2  10   0   0
        L4   0   0   4 117   8   0  16   0   0
        L5   0   0   0   4  75   8   6   2   0
        L6   0   0   0   0   0  15   6   0   0
        L7   0   0   0   0   0   5  69   6   0
        L8   0   0   0   0   0   0   1  93   0
        L9   0   0   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.8411
                 95% CI : (0.8162, 0.8639)
    No Information Rate : 0.1864
    P-Value [Acc > NIR] : < 2.2e-16
```

```
Kappa : 0.8147

Mcnemar's Test P-Value : NA
```

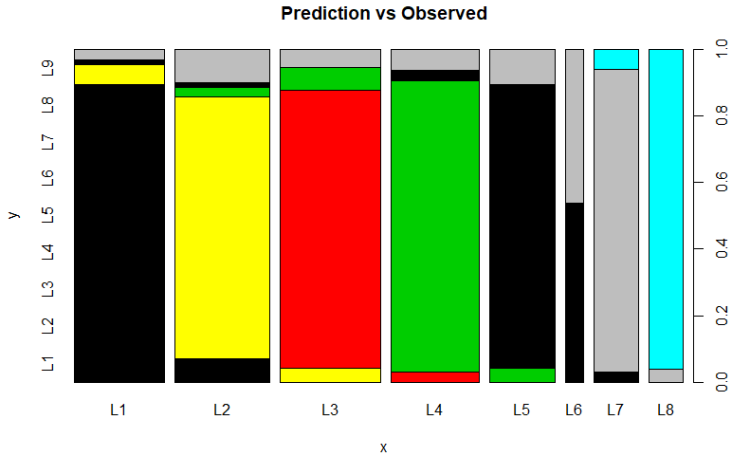
Based on above confusion matrix we can see we have significant number of True positive. In the above matrix, the diagonal number are the number of correctly predicted class and rest of the count in the matrix is incorrect prediction.

Looking at overall statistics, we can see the model as has **accuracy of 84.11%** with confidence interval between 81.62% and 86.39% when evaluated at significance level of 0.05. Also, the 18.64% of no information rate indicate we need bigger and diverse sample to train model with more accuracy. The above prediction was made on dataset used to train model, the accuracy based on above test might not be same of other data sample outside of trainset, to test how well the model perform with dataset outside of training set, we will use it with test set.

Since, sample in test set are completely new to model, predicting class and comparing with observed class with me good way to test the actual accuracy of the model.

```
> test.predict <- predict(expen_ctree, newdata= test.data)
> confusionMatrix(test.predict,test.data$expenses)
Confusion Matrix and Statistics

          Reference
Prediction L1 L2 L3 L4 L5 L6 L7 L8 L9
        L1 59  4  0  0  1  0  2  0  0
        L2  5 55  0  2  1  0  7  0  0
        L3  0  3 62  5  0  0  4  0  0
        L4  0  0  2 57  2  0  4  0  0
        L5  0  0  0  2 38  3  5  0  0
        L6  0  0  0  0  1  6  6  0  0
        L7  0  0  0  0  0  1 30  2  0
        L8  0  0  0  0  0  0  1 24  0
        L9  0  0  0  0  0  0  0  0  0

Overall Statistics

               Accuracy : 0.8401
                 95% CI : (0.8001, 0.8749)
    No Information Rate : 0.1675
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8132

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: L1 Class: L2 Class: L3 Class: L4 Class: L5 Class: L6 Class: L7 Class: L8 Class: L9
Sensitivity            0.9219   0.8871   0.9688   0.8636  0.88372  0.60000  0.50847  0.92308       NA
Specificity            0.9788   0.9548   0.9636   0.9756  0.97151  0.98177  0.99104  0.99728        1
Pos Pred Value         0.8939   0.7857   0.8378   0.8769  0.79167  0.46154  0.90909  0.96000       NA
Neg Pred Value         0.9848   0.9784   0.9938   0.9726  0.98555  0.98950  0.91967  0.99458       NA
Prevalence             0.1624   0.1574   0.1624   0.1675  0.10914  0.02538  0.14975  0.06599        0
Detection Rate         0.1497   0.1396   0.1574   0.1447  0.09645  0.01523  0.07614  0.06091        0
Detection Prevalence   0.1675   0.1777   0.1878   0.1650  0.12183  0.03299  0.08376  0.06345        0
Balanced Accuracy      0.9503   0.9210   0.9662   0.9196  0.92762  0.79089  0.74976  0.96018       NA
```

We can also visualize the correct prediction made by model using plot with color distinguishing class of expenses.



In the above confusion matrix output, we can see the model sustain the accuracy of 84.01%, accuracy of model with train set and test set is almost same with only difference of 0.10%. This signifies that, this model will be able to predict the expense class of any random observation with accuracy between 80.01% and 87.49%. Also, looking at the plot above, we can see the model did bad job in predicting L6 class as 50% of the predictions are incorrect, which is probably because the model was trained with very few samples for L6 class. To improve model accuracy, bigger sample size with proportional distribution can be used.

Now we will plot a scatter plot between age and BMI of smoker on low expenses class and high expense.

```
low.expen <- filter(test.data,test.data$expenses %in% c("L1","L2","L3","L4") | smokeryes == "1")
high.expen <- filter(test.data,test.data$expenses %in% c("L5","L6","L7","L8","L9")| smokeryes == "1")

x1<- ggplot(low.expen, aes(age, bmi)) +
  geom_jitter(color = "blue", alpha = 0.5) +
  theme_light()

y1 <- ggplot(high.expen, aes(age,bmi )) +
  geom_jitter(color = "green", alpha = 0.5) +
  theme_light()

p1 <- plot_grid(x1, y1)
title <- ggdraw() + draw_label("Correlation between Age and bmi of smokers in different expense
level",
                               fontface='bold')
plot_grid(title, p1, ncol=1, rel_heights=c(0.1, 1))
```

In the above plot, we can observe that the concentration of low expense customer is higher toward age below 25 and BMI below 35. And in the other hand, the concentration of high expenses class is more among with age more than 55 and BMI above 25.
Similarly, multiple graphical and statistical test such as t-test and ANOVA test can be performed in dataset to drill down to target demographic based on expenses.

Conclusion

For multiple linear regression method, we have explored the multiple linear regression model and applied it to predict expenses and seen the correlation between predicted and actual results. An important consideration when selecting the model for the multiple linear regression analysis is the model fit. Adding independent variables to a multiple linear regression model will always increase the amount of explained variance in the dependent variable (typically expressed as R^2). As the above comparison of two models, the performance is quite similar so we will keep the new model to the predicting part. The result of the prediction shows that model was able to predict the premium insurance for policyholders with a mean difference of ~8.73%. The size of some of the errors is a bit concerning, but not surprising given the nature of insurance expense data. The errors in the model are close to zero so the model our model is performing fairly well. Since the R-squared value is 0.7503, we know that the model explains more than 75 percent of the variation in

the dependent variable. At the end of this part, two examples of two patients present the use of multiple linear regression model to help the insurance company predict the health insurance.

From the above developed classification model, the predictor columns like age, sex, region, children and BMI determine the value of the "predictor attribute" medical expenses class. In the training set, the predictor attribute/expense class is known. When we train the model, the classification algorithm determined how the value of the predictor attributed was reached, what relationship exists between the predictors and the decision, the structure if this then that rule was developed which we visualized as decision tree.

An insurance company with similar model can develop a model which will help them decide if they should sell insurance to their new client. If they wish to sell the insurance plan, they can evaluate the what level of expenses the customer is likely incur in medical bills, based on that company can decide many aspects such as what factor should they consider in policy, what should be the insurance amount or the monthly premium of the insurance plan. Deciding these crucial questions in business operation can prevents insurance company millions of dollars.

We train our model with 944 observation and tested with different 394 observation; the model has accuracy of 84.01%. However, from above modeling we can conclude that, to improve the model accuracy we need use relatively larger sample with proportional distribution of each class. Since, the decision tree is based on train set, it reflects the decision making based on data used to train model. In our sample, we had small proportion of one of the expenses class, predict of the same class was made with least accuracy. This infers that, not enough data was feed to model to understand the classification more accurately.

References:

Dataset from Kaggle,
https://www.kaggle.com/noordeen/insurance-premium-prediction
K Means Clustering Using R,
https://rpubs.com/abdul_yunus/Kmeans_Clustering
A Guide to Machine Learning in R for Beginners: Decision Trees,
https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb
Rstudio, pubs - statistic,
http://rstudio-pubs-static.s3.amazonaws.com/254185_676cb9ff7fee4ab9b8d805cb7ebe92c4.html#
Badal Kumar, Acadgild,
https://acadgild.com/blog/premium-insurance-policyholders-using-linear-regression-with-r