Foundations of Artificial Intelligence (ISTM 6214)

Insurance Charges Analysis Report

Group 7

Yubika Bhattarai, Christina Huynh, Ankita Midha

December 12, 2021

Original Work Statement

We, the undersigned, certify that the actual composition of this proposal was done by us and is original work.

Full Name	Signature
Yubika Bhattarai	Yubika Bhattarai
Christina Huynh	Christina Huynh
Ankita Midha	Amidha

Table of Content

1. Executive Summary

Insurance charges are usually unstable throughout the life of customers. The report did extra or less for their insurance premium charges each year based on customers' medical histories and demands. Insurance companies often have a hard time providing clear explanations to their customers about insurance charges and benefits. Therefore, customers might not be satisfied with their charges and the insurer. This can lead to a deduction of loyal customers and sales to insurance companies.

We have learned different skill sets in data analytics and methodologies throughout the course. Many organizations have applied these solutions to predict customers' behaviors, financial trends, or insurance calculations. Therefore, we would like to use this insurance situation to analyze further the factors that are significant to the change of insurance charges. We also aim to find the best methodology to interpret insurance data. This report can be helpful for insurance companies, especially data scientists, to investigate the factors that affect customers' insurance charges yearly.

The objectives of this report include:

- Use Linear Regression and Logistic Regression to find the most significant factor that impacts insurance charges.
- Execute k-NN model to estimate customer's behavior if they are likely a smoker or not.
- Run Decision Tree to comprehend the significance of each factor on insurance charge by displaying on a tree view.

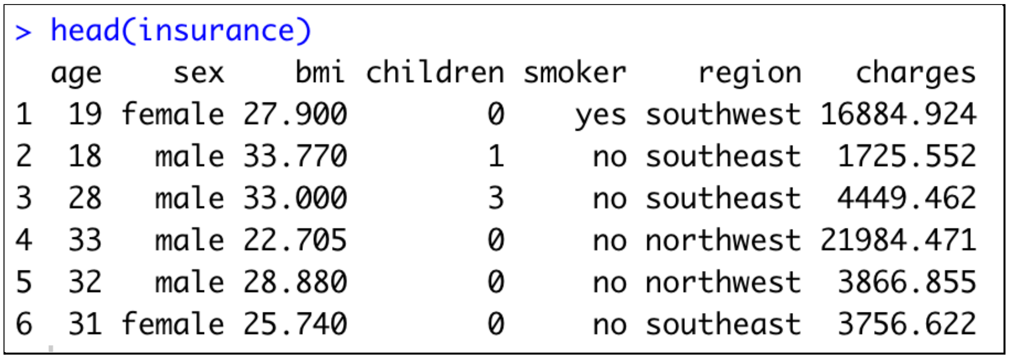
2. Data Description

The data was sourced from Kaggle (https://www.kaggle.com/mirichoi0218/insurance) by Brett Lanz in Machine Learning with R book. This dataset has 7 columns and 1339 rows, showing the factors to predict insurance charges. These factors are both numerical and non-numerical and are commonly used by insurance companies. The following table describes the details of each variable:

Age	The age of customers who wish to have insurance that ranges from 18 to 64 (numerical)
Sex	The gender of customers: Female and Male (binary)
BMI	Body mass index is calculated by a person's weight (kg) divided by the square of height (m^2). A good BMI is usually from 18.5 to 24.9 (numerical)
Children	The number of children/Number of dependents. It is from 0 to 5 in this dataset (numerical)
Smoker	Whether the customer is a smoker (binary)
Region	The region that the customer currently lives in the U.S. (categorical)
Charges	Insurance charge billed by insurance companies (numerical)

Table 1: Data Description

This dataset contains many factors that would help us analyze each factor's effect on insurance charges. In order to understand the snapshot of the dataset, we use head() function to see the first part of the dataset:



```
> head(insurance)
  age    sex    bmi children smoker    region   charges
1  19 female 27.900        0    yes southwest 16884.924
2  18   male 33.770        1     no southeast  1725.552
3  28   male 33.000        3     no southeast  4449.462
4  33   male 22.705        0     no northwest 21984.471
5  32   male 28.880        0     no northwest  3866.855
6  31 female 25.740        0     no southeast  3756.622
```

Figure 1: First part of the dataset

3. Research Questions

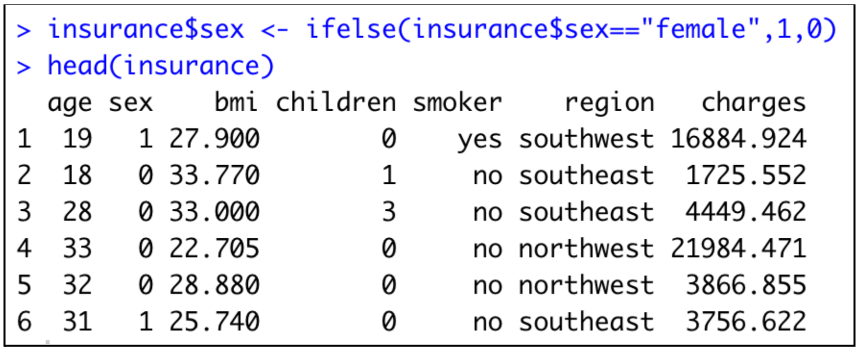
Nowadays, many insurance companies are trying to offer the best rate for consumers based on their demands. However, it is not easy for companies to issue a reasonable rate to satisfy consumers. In order to provide a reasonable insurance rate with many benefits, insurance companies usually perform surveys and analyses frequently. This method helps companies to understand the trend and demands of their customers. The factors commonly used to define an insurance rate include age, sex, employment, number of dependents, regions, etc. Companies also use customers' medical histories to finalize insurance quotes. It is not surprising that insurance premiums have increased yearly, and this is because every individual and family has different purposes and conditions every year or month. Customers' backgrounds, such as health conditions, salaries, number of dependents, quickly change in a short period.

Nevertheless, some consumers do not comprehend why their insurance charges increase. This leads to dissatisfaction of consumers with insurance companies. As a result, insurance companies could lose loyal customers. Furthermore, customers could not take advantage of insurance benefits. Hence, this analysis will use the above dataset from Kaggle and different methodologies including linear regression, logistic regression, k-NN, and decision tree to identify:

- The factors affecting insurance charges and which factor has the most significant effect on insurance charges using Linear Regression Model
- Whether smoker or non-smoker increases the insurance charges
- Whether a smoker or non-smoker can be classified using Logistic Regression and which factors had the most significant effect on smoking
- The best k from the k-NN model to predict a customer is a smoker or not
- The most effective methodology to use for insurance charges prediction

4. Data Preparation

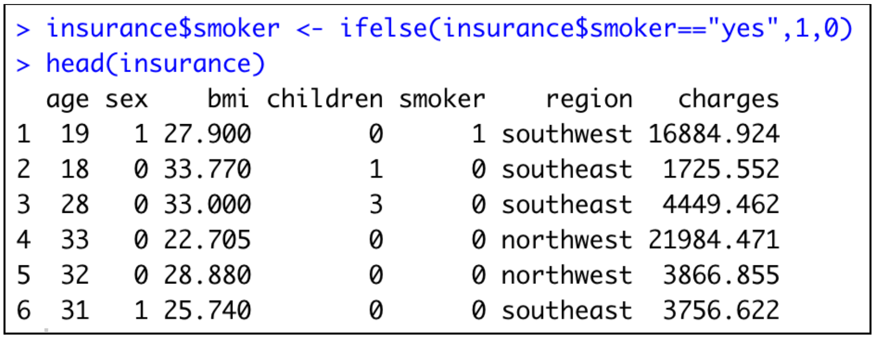
Since sex and smoker are binary types, we will convert them to dummy variables 1 and 0. Using the ifelse() function, females are represented as 1, and males are represented as 0.



```
> insurance$sex <- ifelse(insurance$sex=="female",1,0)
> head(insurance)
  age sex    bmi children smoker    region   charges
1  19   1 27.900        0    yes southwest 16884.924
2  18   0 33.770        1     no southeast  1725.552
3  28   0 33.000        3     no southeast  4449.462
4  33   0 22.705        0     no northwest 21984.471
5  32   0 28.880        0     no northwest  3866.855
6  31   1 25.740        0     no southeast  3756.622
```

Figure 2: Sex Column: Female = 1, Male = 0

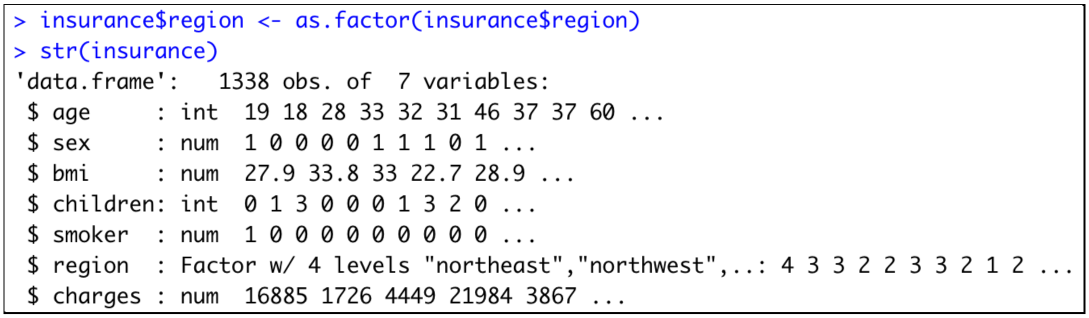
Next, smoker type is also converted to dummy variables as yes = 1 and no = 0, the result is shown below:



```
> insurance$smoker <- ifelse(insurance$smoker=="yes",1,0)
> head(insurance)
  age sex    bmi children smoker    region   charges
1  19   1 27.900        0      1 southwest 16884.924
2  18   0 33.770        1      0 southeast  1725.552
3  28   0 33.000        3      0 southeast  4449.462
4  33   0 22.705        0      0 northwest 21984.471
5  32   0 28.880        0      0 northwest  3866.855
6  31   1 25.740        0      0 southeast  3756.622
```

Figure 3: Smoker Column: Yes = 1, No = 0

Last but not least, we will convert the Region column into a categorical variable. Using the as.factor() function, we successfully converted the Region column into factor types from 1 to 4, representing the four regions: northeast, northwest, southeast, and southwest. Below is the summary of column variable type of this dataset after data preparation:



```
> insurance$region <- as.factor(insurance$region)
> str(insurance)
'data.frame':	1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : num  1 0 0 0 0 1 1 1 0 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

Figure 4: Type of each column

5. Data Exploration

5.1. Histogram

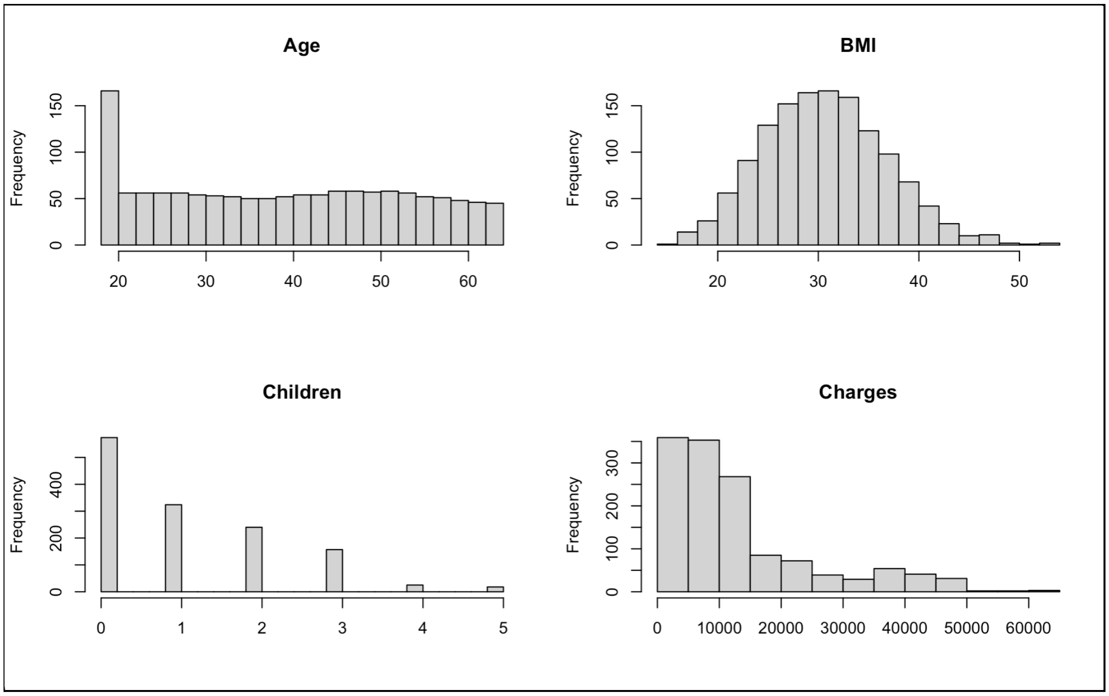


Figure 5: Histograms of Age, BMI, Children, and Charges

After preparing the dataset, we created some diagrams to see how each factor contributes to insurance charges. Age, BMI, and the number of children could be highly significant to the insurance charges. Because BMI is the body mass measurement usually used in the medical field, this factor could directly affect insurance charges. Moreover, different age ranges would have different medical needs and issues. Hence, older people usually have a higher insurance rate with more benefits such as hospitalization, prescriptions, and medical equipment. Last but not least, the number of children can be counted as the number of additional costs to parents such as prenatal care, child support, etc. Therefore, parents might need additional benefits to satisfy their demands.

5.2. Boxplot

We also executed different boxplots for each factor to its outlier. This way can help us to analyze our dataset quickly.

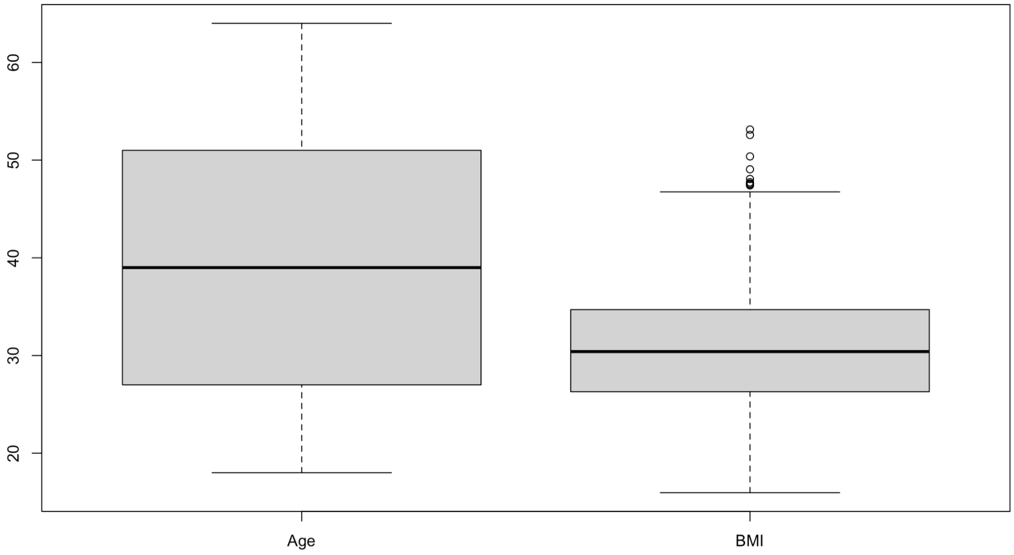


Figure 5.1: Boxplot between Age and BMI

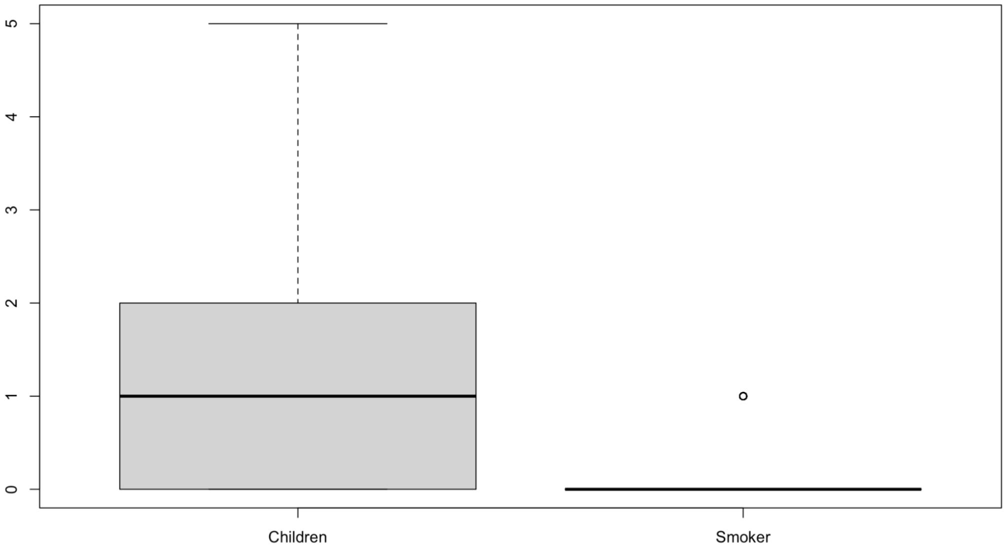


Figure 5.2: Boxplot between Children and Smoker

Since we have converted the Smoker column into dummy variables, the outlier is displayed differently compared to other factors. We can see from the boxplots above that Smoker and BMI have outliers. The number of children is small in this dataset; so, the number of children might have a lower significance than other factors in the insurance charges.

With basic knowledge and analysis, our initial thought is that Age, BMI, and the number of children would significantly affect the insurance premium charges of individuals and families. In order to find out which factor has the highest significance to insurance charges. Advanced methodologies will be executed in the report.

6. Methodology

6.1. Linear Regression

Linear Regression is the simplest form of predictive modeling and is a good starting point to understanding the data set. It is used to predict the value of a variable based on the value of another variable. There are advantages of linear regression, like uncovering patterns and relationships scientifically; however, there is a significant limitation of linear regression, which is the apparent correlation between dependent and predictor variables which could be misleading at times. With our data set, we aim to fit a linear regression model to predict the insurance premium (charges) using the predictor variables.

Exploratory Data Analysis

Before building the linear regression model, we conducted exploratory data analysis to find the correlation between insurance premium (charges) and other independent variables.

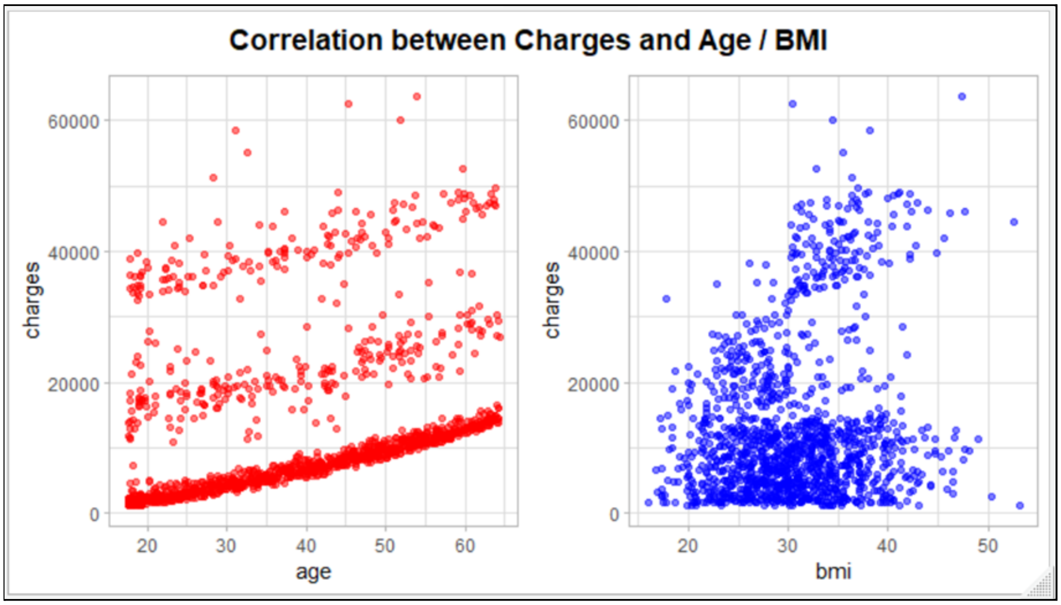


Figure 6:Correlation Matrix Charges with Age and BMI

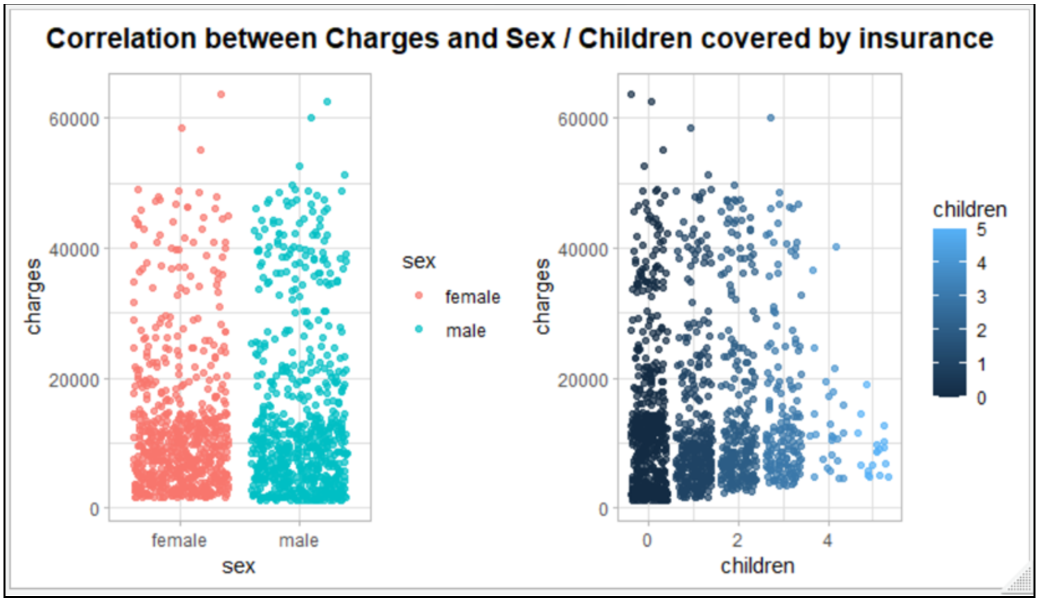


Figure 6.1: Correlation Matrix Charges with Gender and Children

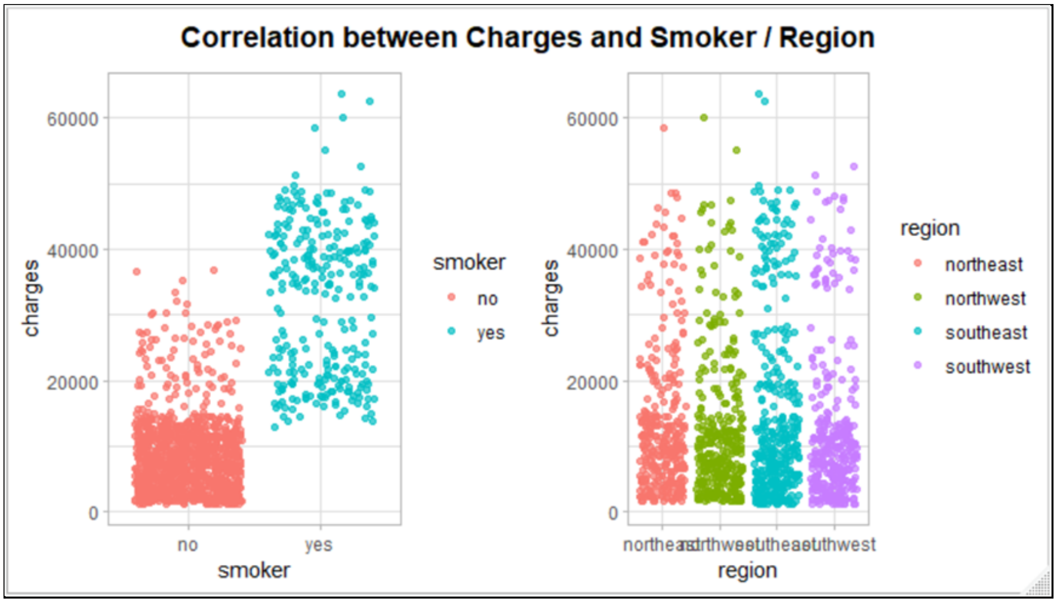


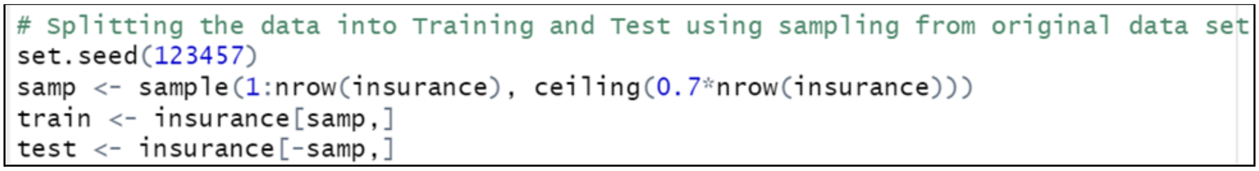
Figure 6.2: Correlation Matrix Charges with Smoker and Region

Observations

- As Age and BMI go up, charges for health insurance also trend upwards
- Charges for insurance with 4-5 children are less than others in the range
- Charges for insurance for Smokers are higher than for Non-smokers
- All the regions are at par with each other in influencing insurance charges

Preparing the Dataset

We split the insurance data into Training (70%) and Test (30%) of the samples

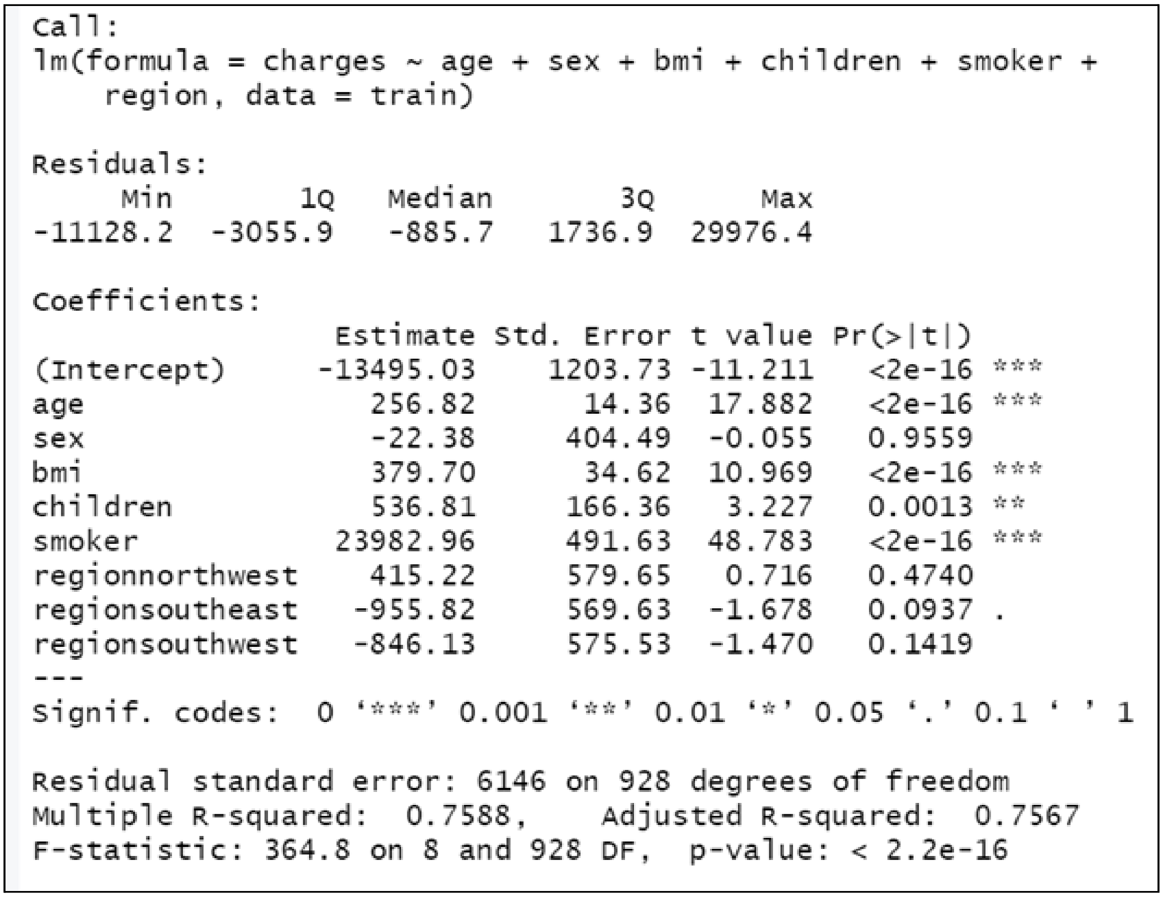


```
# Splitting the data into Training and Test using sampling from original data set
set.seed(123457)
samp <- sample(1:nrow(insurance), ceiling(0.7*nrow(insurance)))
train <- insurance[samp,]
test <- insurance[-samp,]
```

Figure 6.3: R Codes

Building the Linear Regression Model

We built a multiple linear regression model using all predictor variables and arrived at the following summary:



```
Call:
lm(formula = charges ~ age + sex + bmi + children + smoker + 
    region, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-11128.2  -3055.9   -885.7   1736.9  29976.4 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -13495.03    1203.73 -11.211   <2e-16 ***
age               256.82      14.36  17.882   <2e-16 ***
sex               -22.38     404.49  -0.055   0.9559    
bmi               379.70      34.62  10.969   <2e-16 ***
children          536.81     166.36   3.227   0.0013 ** 
smoker          23982.96     491.63  48.783   <2e-16 ***
regionnorthwest   415.22     579.65   0.716   0.4740    
regionsoutheast  -955.82     569.63  -1.678   0.0937 .  
regionsouthwest  -846.13     575.53  -1.470   0.1419    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6146 on 928 degrees of freedom
Multiple R-squared:  0.7588,	Adjusted R-squared:  0.7567 
F-statistic: 364.8 on 8 and 928 DF,  p-value: < 2.2e-16
```

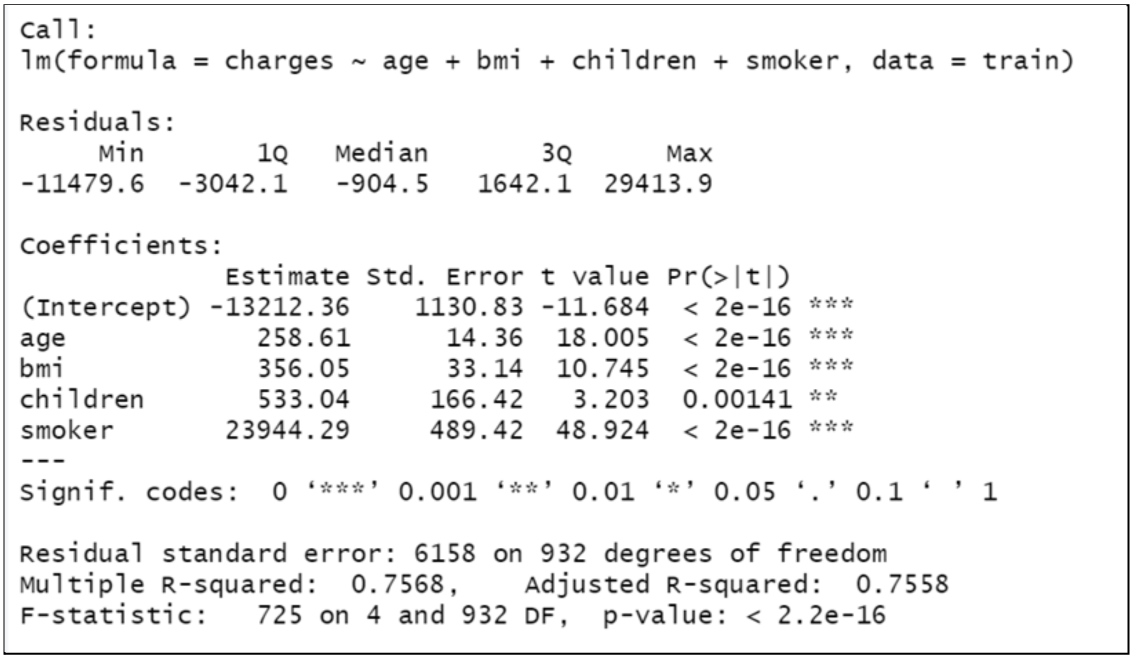
Figure 6.4: Result of lm() function

We observe that the independent variable’s age, BMI, children, and smoker except for gender, are statistically significant with p values less than 0.05 and positively correlated to insurance premium charges.

Performance of the Model

Adjusted R Squared: 0.757 - that implies 75% of the variation of charges could be explained by the set of predictor variables we have included. And, RMSE of this model is 5907.164

We further trained the model without non-significant variables and checked if the performance could be improved. We included age, BMI, children, and smoker predictor variables for this model. Figure 6.5 provides the summary of the model.



```
Call:
lm(formula = charges ~ age + bmi + children + smoker, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11479.6  -3042.1   -904.5   1642.1  29413.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13212.36    1130.83 -11.684  < 2e-16 ***
age            258.61      14.36  18.005  < 2e-16 ***
bmi            356.05      33.14  10.745  < 2e-16 ***
children       533.04     166.42   3.203  0.00141 **
smoker       23944.29     489.42  48.924  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6158 on 932 degrees of freedom
Multiple R-squared:  0.7568,    Adjusted R-squared:  0.7558
F-statistic:   725 on 4 and 932 DF,  p-value: < 2.2e-16
```

Figure 6.5: Result of lm() function

The Adjusted R-squared for this model is 0.7558, which is not different from the previous model, while the RMSE was 5868.713, slightly less than the previous model. Overall, the performance is similar between the two models; however, the last model applies only the significant variables affecting the insurance premiums.

From the Linear Regression Analysis, we find:

- Age, BMI, number of children, and smoking are the ones that drive the insurance premium charges.

- Smoking seems to have the most influence on the insurance premium charges

- The RMSE of the model is relative with the above machine learning modeling techniques high and not fit for prediction accuracy

6.2. Logistic Regression

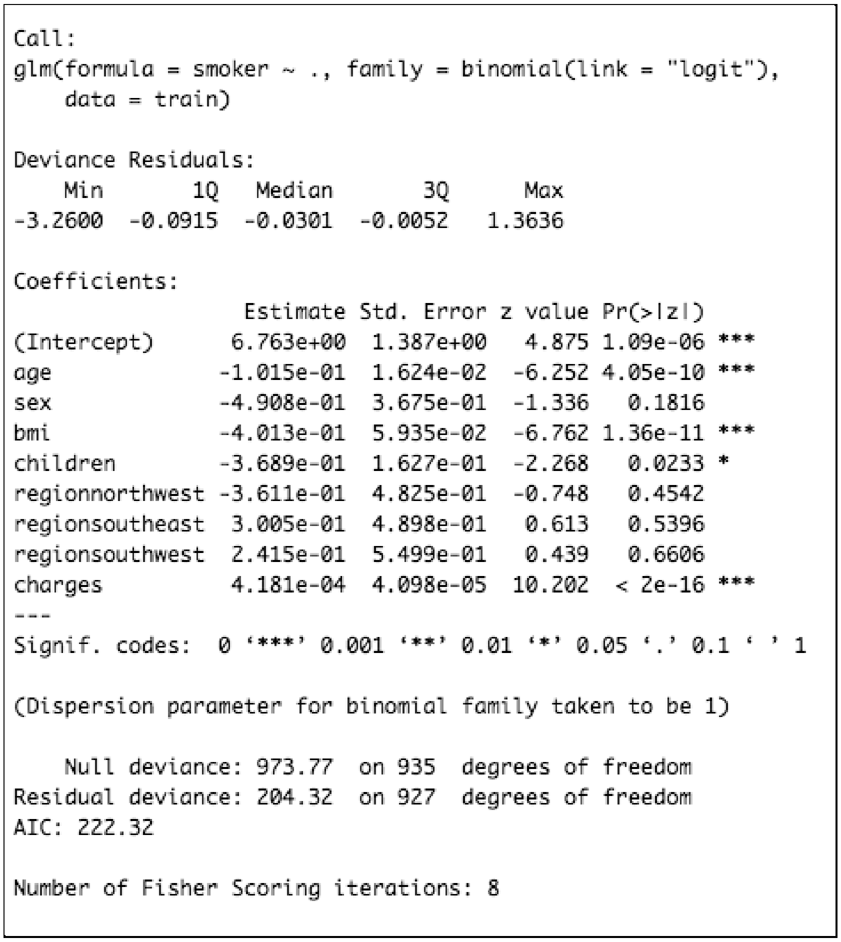
Logistic regression was used to classify if a customer is a smoker or non-smoker. Moreover, it was also used to understand which factors have the most substantial influence in predicting whether someone is a smoker or non-smoker.

Preparing the dataset

To create our logistic regression model, we split the data into two sets: training (70%) and validation (30%). The Logistic Regression model was created using the training dataset, and the validation dataset was used to make predictions.

Building the Logistic Regression Model

Figure 6.6 is the output of the Logistic Regression algorithm. It suggests that age, BMI, and charges have the most decisive influence on smoking. This conclusion was formed because those three factors have the most significant coefficients and strong p-values, and a strong p-value supports those three variables are statistically significant. However, because insurance charge is an output and a central variable we are trying to predict in this report, it will not be significant when analyzing smoking.

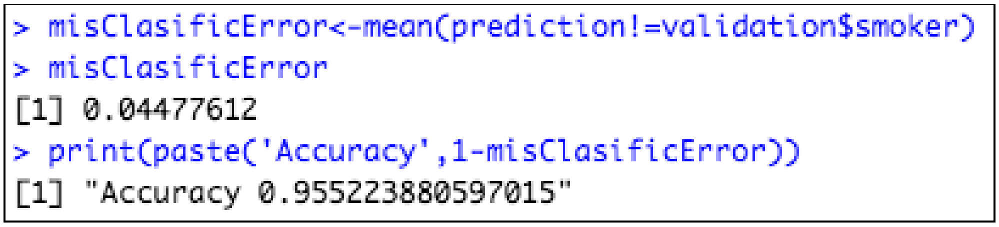


```
Call:
glm(formula = smoker ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2600  -0.0915  -0.0301  -0.0052   1.3636

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.763e+00  1.387e+00   4.875 1.09e-06 ***
age             -1.015e-01  1.624e-02  -6.252 4.05e-10 ***
sex             -4.908e-01  3.675e-01  -1.336   0.1816
bmi             -4.013e-01  5.935e-02  -6.762 1.36e-11 ***
children        -3.689e-01  1.627e-01  -2.268   0.0233 *
regionnorthwest -3.611e-01  4.825e-01  -0.748   0.4542
regionsoutheast  3.005e-01  4.898e-01   0.613   0.5396
regionsouthwest  2.415e-01  5.499e-01   0.439   0.6606
charges          4.181e-04  4.098e-05  10.202  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 973.77  on 935  degrees of freedom
Residual deviance: 204.32  on 927  degrees of freedom
AIC: 222.32

Number of Fisher Scoring iterations: 8
```

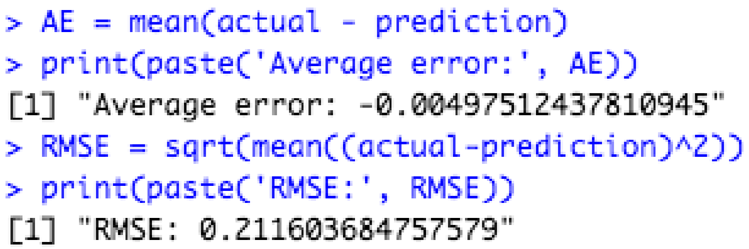
Figure 6.6: Logistic Regression Output

Performance of the Model

The accuracy of our prediction was found to be 0.96, which is a strong number, suggesting that our model can accurately predict if the customer is a smoker or non-smoker. Furthermore, our AE was approximately 0, and the RMSE was 0.21.



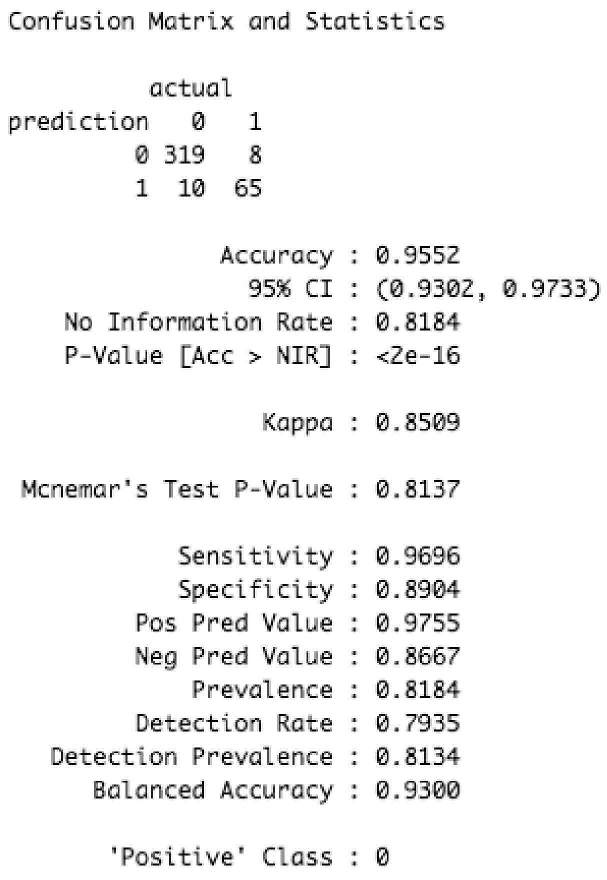
```
> misClasificError<-mean(prediction!=validation$smoker)
> misClasificError
[1] 0.04477612
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.955223880597015"
```



```
> AE = mean(actual - prediction)
> print(paste('Average error:', AE))
[1] "Average error: -0.00497512437810945"
> RMSE = sqrt(mean((actual-prediction)^2))
> print(paste('RMSE:', RMSE))
[1] "RMSE: 0.211603684757579"
```

Figure 6.7: Model Accuracy and Error Outputs of Logistic Regression Model

The confusion matrix below (Figure 6.8) further indicates that our model accurately identifies true positives and negatives. True positives are individuals who were predicted to be smokers and were found to be actual smokers in the validation data. The figure below shows that 65 out of 75 classifications were true positives. True negatives are individuals classified as non-smokers by our model who were actual non-smoker in the validation dataset. The number of individuals correctly identified as true negatives is 319 out of 327. Using information from the confusion matrix, we were able to calculate the sensitivity and specificity of our model. Sensitivity is the true positive rate of our model and was calculated to be 0.97. Sensitivity is calculated by dividing true positives by the sum of both true positives and false negatives. The specificity of our model is 0.89. The specificity identified the proportion of true negatives and was calculated by dividing true negatives by the sum of true negative and false positive. The specificity and sensitivity rates suggest that our current model might be better at predicting positives than predicting negatives, in essence, identifying if someone is a smoker rather than a non-smoker.



```
Confusion Matrix and Statistics

          actual
prediction   0   1
         0 319   8
         1  10  65

               Accuracy : 0.9552
                 95% CI : (0.9302, 0.9733)
    No Information Rate : 0.8184
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8509

 Mcnemar's Test P-Value : 0.8137

            Sensitivity : 0.9696
            Specificity : 0.8904
         Pos Pred Value : 0.9755
         Neg Pred Value : 0.8667
             Prevalence : 0.8184
         Detection Rate : 0.7935
   Detection Prevalence : 0.8134
      Balanced Accuracy : 0.9300

       'Positive' Class : 0
```

Figure 6.8: Confusion Matrix, Sensitivity and Specificity Output for Logistic Regression

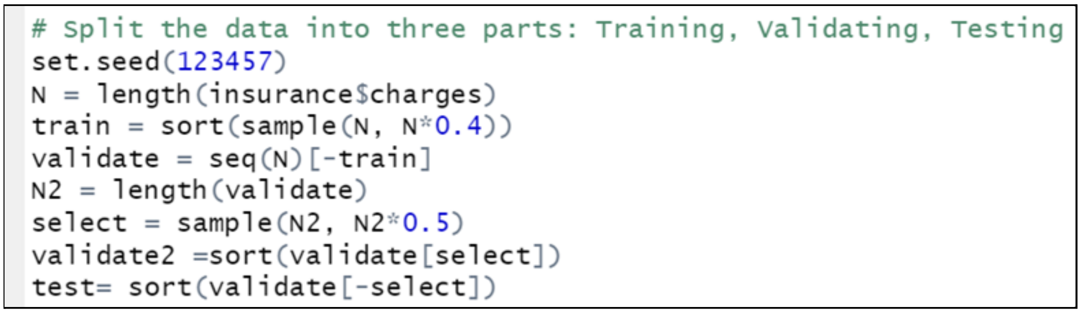
The information provided by the logistic regression model could be valuable in understanding insurance charges because it can reasonably accurately identify a smoker. Smoking habits could be a potential factor in determining insurance charges because insurers are known to place a premium on unhealthy habits such as smoking.

6.3. k-NN Model

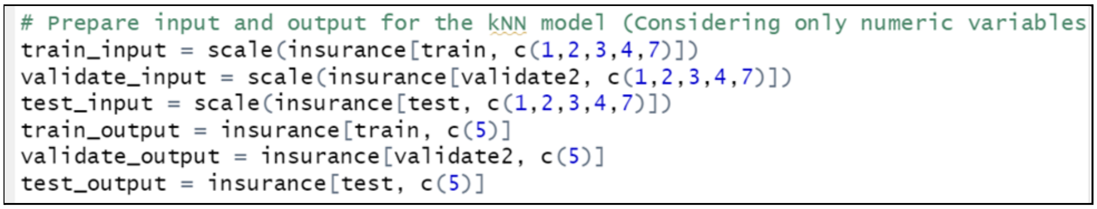
We also leveraged the k-NN model to find the best nearest neighbor to predict whether a customer will be a smoker or non-smoker using certain features of age, sex, BMI, children, and insurance charges. The k-NN uses the counting of the majority of votes from its nearest neighbors for classification models.

Normalization and Splitting of the Dataset

Our dataset has different units of measurements, such as the BMI and the insurance charges. In k-NN, we measure the distances between pairs of samples, which are also influenced by the measurement units. To avoid this misclassification, we normalized our feature variables in Figure 6.9. We used the training set to train the k-NN model, the validation set to select the optimal K, and the testing set to evaluate the model performance.



```
# Split the data into three parts: Training, Validating, Testing
set.seed(123457)
N = length(insurance$charges)
train = sort(sample(N, N*0.4))
validate = seq(N)[-train]
N2 = length(validate)
select = sample(N2, N2*0.5)
validate2 =sort(validate[select])
test= sort(validate[-select])
```

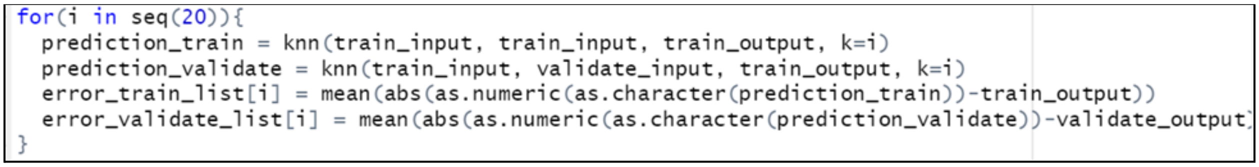


```
# Prepare input and output for the kNN model (Considering only numeric variables
train_input = scale(insurance[train, c(1,2,3,4,7)])
validate_input = scale(insurance[validate2, c(1,2,3,4,7)])
test_input = scale(insurance[test, c(1,2,3,4,7)])
train_output = insurance[train, c(5)]
validate_output = insurance[validate2, c(5)]
test_output = insurance[test, c(5)]
```

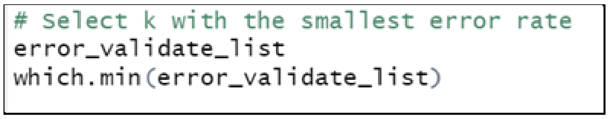
Figure 6.9: Data Sample split and k-NN Model preparation

Find the Best k

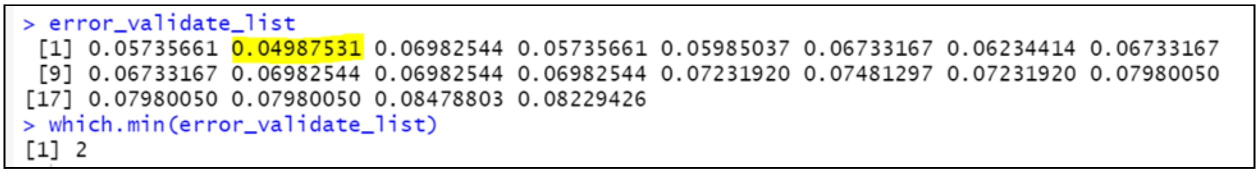
We trained the model with k in the range of 1 to 20 using the training set and tested it on the prediction set.



```
for(i in seq(20)){
  prediction_train = knn(train_input, train_input, train_output, k=i)
  prediction_validate = knn(train_input, validate_input, train_output, k=i)
  error_train_list[i] = mean(abs(as.numeric(as.character(prediction_train))-train_output))
  error_validate_list[i] = mean(abs(as.numeric(as.character(prediction_validate))-validate_output
}
```



```
# Select k with the smallest error rate
error_validate_list
which.min(error_validate_list)
```



```
> error_validate_list
 [1] 0.05735661 0.04987531 0.06982544 0.05735661 0.05985037 0.06733167 0.06234414 0.06733167
 [9] 0.06733167 0.06982544 0.06982544 0.06982544 0.07231920 0.07481297 0.07231920 0.07980050
[17] 0.07980050 0.07980050 0.08478803 0.08229426
> which.min(error_validate_list)
[1] 2
```

Figure 6.10:Best k from the Model

As shown in Figure 6.11 below, the validation error initially increases when k=2 and then vacillates in a stable state.

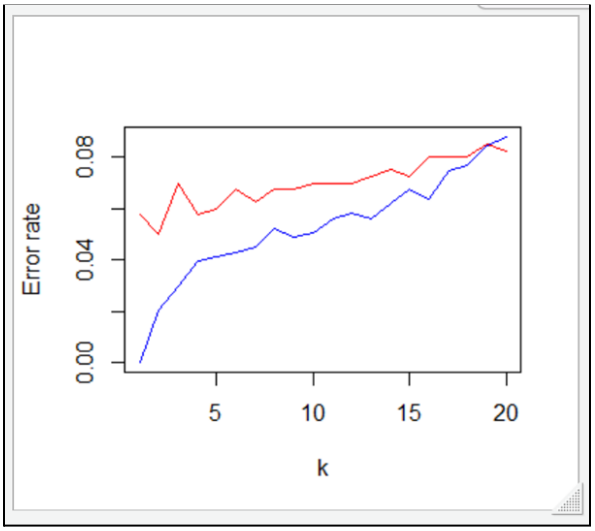
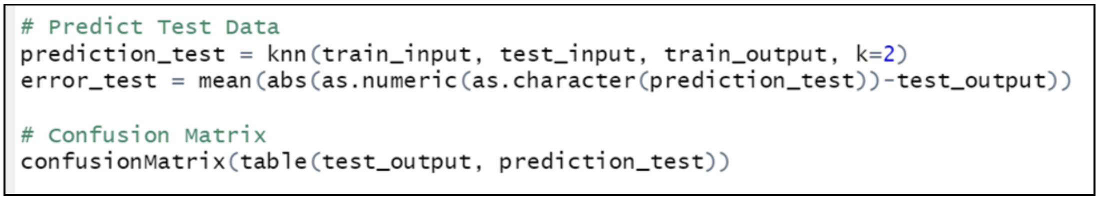
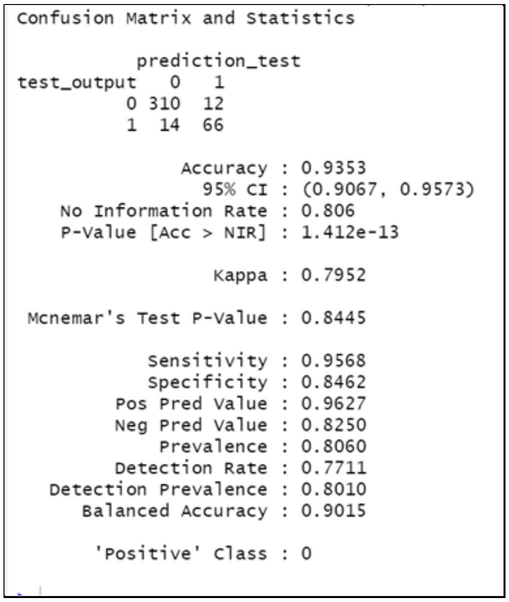


Figure 6.11: Error Rate Plot

Model Prediction with best k=2



```
# Predict Test Data
prediction_test = knn(train_input, test_input, train_output, k=2)
error_test = mean(abs(as.numeric(as.character(prediction_test))-test_output))

# Confusion Matrix
confusionMatrix(table(test_output, prediction_test))
```



```
Confusion Matrix and Statistics

           prediction_test
test_output   0   1
          0 310  12
          1  14  66

               Accuracy : 0.9353
                 95% CI : (0.9067, 0.9573)
    No Information Rate : 0.806
    P-Value [Acc > NIR] : 1.412e-13

                  Kappa : 0.7952

 Mcnemar's Test P-Value : 0.8445

            Sensitivity : 0.9568
            Specificity : 0.8462
         Pos Pred Value : 0.9627
         Neg Pred Value : 0.8250
             Prevalence : 0.8060
         Detection Rate : 0.7711
   Detection Prevalence : 0.8010
      Balanced Accuracy : 0.9015

       'Positive' Class : 0
```

Figure 6.12:Prediction Accuracy

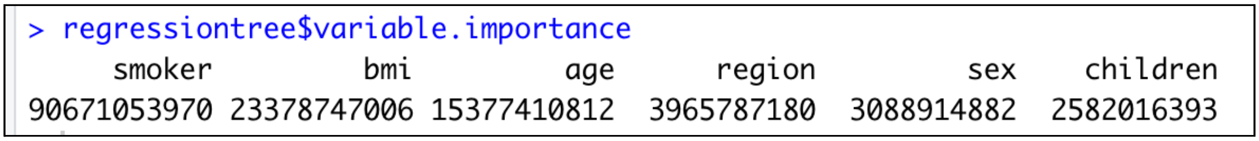
The best k=2 gives testing accuracy of 93.5%, which is good, and sensitivity of 95.6%, while specificity is 84.6%. Given reasonable specificity and sensitivity, the model can accurately predict the true positives (Smokers) and True Negative (Non- Smokers).

6.4. Decision Tree

A Decision Tree is an exploratory tool used to explain and interpret data in a tree figure. The purpose of a decision tree is to represent the data more accurately and efficiently. Therefore, many data analytics have been applying this tool to analyze a large amount of data and make better decisions.

We have built a decision tree with minsplit = 10, minbucket = 7, and a complexity parameter of 0.001. This gives a detailed view of the decision tree. Using the print() function to display the decision tree, a smoker has the highest chance to be the main factor that affects insurance charges (Appendix 1).

We printed the critical variables to see better whether or not other factors are significant to insurance charges.



```
> regressiontree$variable.importance
     smoker         bmi         age      region         sex    children
90671053970 23378747006 15377410812  3965787180  3088914882  2582016393
```

Figure 6.13: Important Variables

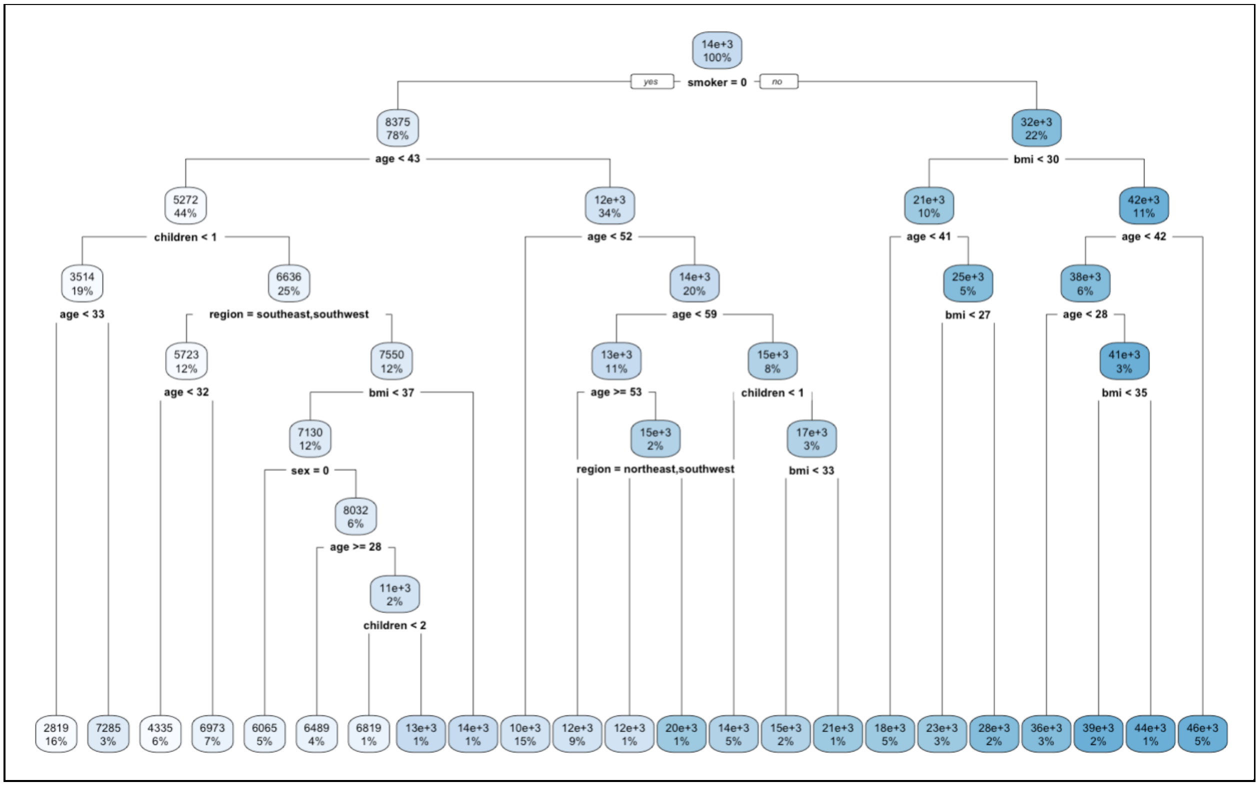


Figure 6.14: Decision Tree (minsplit=10,minbucket=7,cp=0.001)

Figure 6.14 indicates that non-smokers have less chance of increasing insurance charges. Whereas smokers have a higher chance to receive higher insurance charges. To better understand the decision tree, we executed another decision tree with lower minsplit and minbucket (minsplit=5,minbucket=5, maxdepth=10).

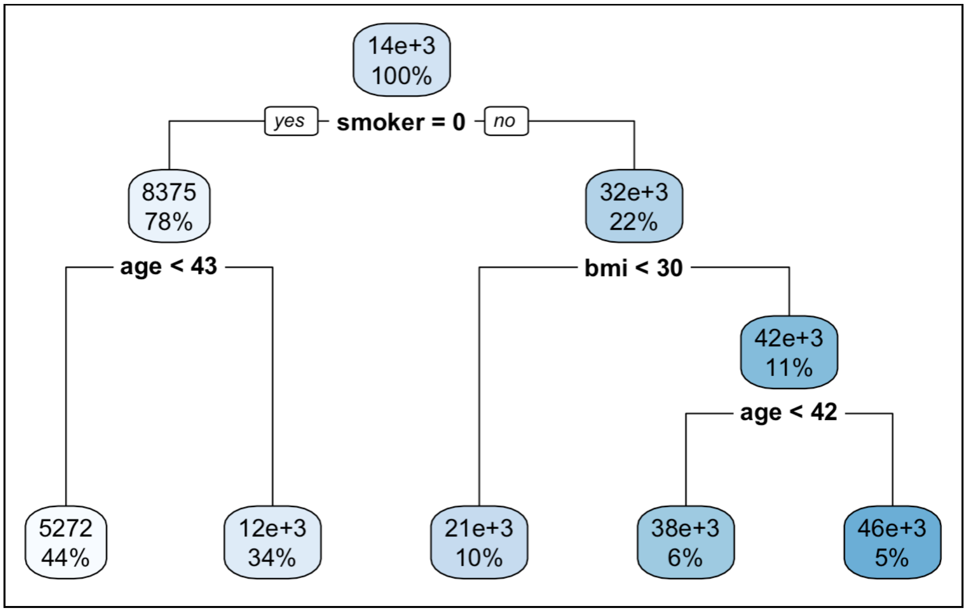
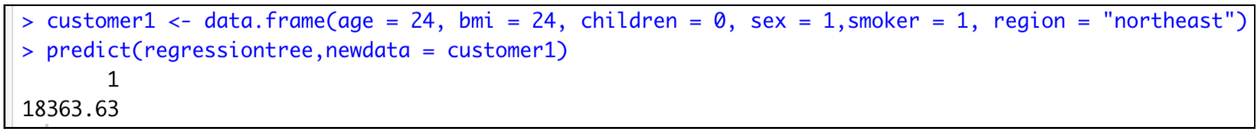


Figure 6.15: Decision Tree (minsplit=5,minbucket=5, maxdepth=10)

Next, we will predict insurance charges for 3 different types of customers. It will help us clearly understand what smokers under 43 years old and who have low BMI (under 25) might get for their insurance premium charges.

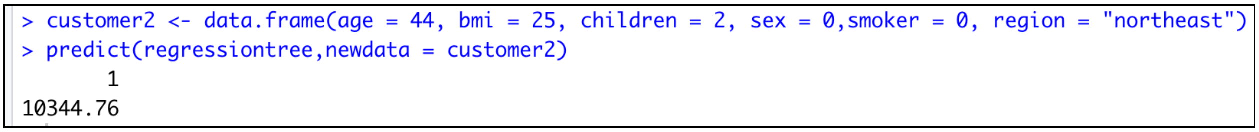
Customer 1: Age = 24, sex = 1, BMI = 24, Children = 0, smoker = 1, region = northeast



```
> customer1 <- data.frame(age = 24, bmi = 24, children = 0, sex = 1,smoker = 1, region = "northeast")
> predict(regressiontree,newdata = customer1)
       1
18363.63
```

Figure 6.16: Insurance Charge of Customer 1

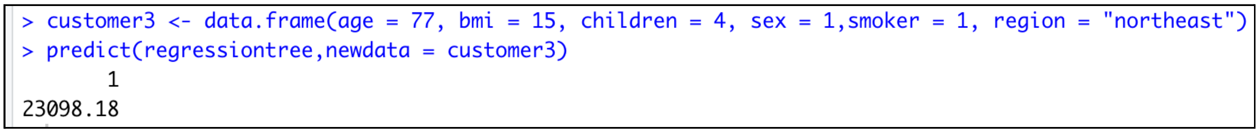
Customer 2: Age = 44, sex = 0, BMI = 25, children = 2, smoker = 0, region = northeast



```
> customer2 <- data.frame(age = 44, bmi = 25, children = 2, sex = 0,smoker = 0, region = "northeast")
> predict(regressiontree,newdata = customer2)
       1
10344.76
```

Figure 6.17: Insurance Charge of Customer 2

Customer 3: Age = 77, sex = 1, BMI = 15, children = 4, smoker = 1, region = northeast



```
> customer3 <- data.frame(age = 77, bmi = 15, children = 4, sex = 1,smoker = 1, region = "northeast")
> predict(regressiontree,newdata = customer3)
       1
23098.18
```

Figure 6.18: Insurance Charge of Customer 3

From figures 6.16 and 6.18, we can see that if the customer is a smoker, he/she has insurance charge will be high will have higher insurance the insurance charge will be high if a smoker is an elder.

7. Results and Finding

We can arrive at our research questions for the medical insurance premium data set with the above machine learning modeling techniques. Our findings were as follows:

- **Predictor variables affecting the insurance premium**: With the Linear Regression model, we analyzed that independent variables of age, BMI, children, smoker have the most significant relationship with an increase in insurance premiums for any customer. Furthermore, by analyzing the coefficients in our linear regression model and the p-value, we determined which factors had the most remarkable and statistically significant impact. The factor we found to have the most significant effect on insurance charge was smoking, and both our decision tree model and linear regression supports this finding.

- **Predict if a customer is a smoker or non-smoker given a set of predictor variables (characteristics) of a customer**: With the k-NN modeling technique, we arrived at a 93.5 % model accuracy with 95.6% sensitivity and 84.6% specificity with the ability of the model to predict both smokers and non-smokers. Our logistic regression models also had similar predictive performances and results. Our Logistic regression model had an accuracy of 95.5%, a sensitivity of 97%, and a specificity of 89%. With Logistic regression, we can further narrow down which factors had the most significant impact on smoking and found those factors to be age and BMI. Both our k-NN and Logistic models accurately predicted smokers and non-smoker.

8. Model Comparison

Model comparison is an excellent tool for analyzing the predictive performances of models. By comparing our models, we can understand if a single model has the best predictive power or if a multiple model approach is better for confirmation and reproducibility.

During our data exploration, we created correlation matrices to understand which factors might have the most significant impact on insurance charges. We noticed that an increase in the value of variables such as BMI and age seem to have a positive correlation with charges. Another variable that stood out to us during our data exploration was smoking, and smoking seemed to have a significant impact on charges and was a variable of great interest to us. Once we built our linear regression model, observations similar to the matrices from our data exploration were found. In the linear regression model, the coefficients for BMI, age, and smoking are approximate: 380, 250, 24000, respectively, while all other coefficients were much less significant in value. The three variables mentioned above also had a p-value < 0.05 and were thus statistically significant.

To further understand the variable, smoking, we tried to predict whether a customer is a smoker or a non-smoker using Logistic Regression. Our Logistic Regression model performed relatively well and had an accuracy of 0.96, an average error of approximately 0, and a root mean square error of 0.21. The model had a sensitivity of 0.97 and a specificity of 0.89. The Logistic Regression Model was able to identify a smoker and a non-smoker reasonably accurately. To compare our findings from the Logistic Regression Model, we built a k-NN model that also tried to predict if a customer is a smoker using nearest neighbors. When the k-NN model was built using k= 2, it gave an accuracy of 0.94, a sensitivity of 0.97, and a specificity of 0.85. By comparing the findings of the two models, k-NN and Logistic, we ensured that the model results were reproducible. We found that both k-NN and Logistic models had similar predictive performances. Our findings suggest that the strength of the logistic regression model is that it could tell us which factors have the most significant influence on smoking. The Logistic model also had a slightly better accuracy of 95.5%, compared to 93.5% for the k-NN model. However, both k-NN and Logistic regression models could be used to predict smoking. Ideally, both models could validate the others' predictions, as was done in this report.

The decision tree model also supports that smokers pay a higher insurance premium. We used the decision tree model to predict the insurance charges for various customers. When comparing the output of the models, we found that the single most significant factor influencing insurance charges is smoking. We know that BMI positively correlates with insurance charges from our linear regression matrices. However, the decision tree model suggests that smoking on an insurance charge is so significant that a customer with a healthy BMI could pay a hefty premium when they are a smoker (Figure 6.17). To further understand the predictive power of our decision tree, we analyzed three customers. Customer three is a smoker with a low BMI and has a predicted insurance charge of 23098. Customer 1, a smoker, has a predicted insurance charge of 18363. Customer 2, a non-smoker, has a predicted insurance charge of 10344. According to our decision tree model, customer 2, a non-smoker, is predicted to pay a significantly lower premium.

Overall, the model comparison suggests that each model has its unique strengths and weaknesses. When we compared k-NN to logistic regression, we found both models with similar output and predictive performances. One strength of the logistic model was that we could analyze individual variables to see the most significant impact. The decision tree model and linear regression also shared similar predictions, such as the positive correlation between smoking and insurance charges. By using multiple predictive models and comparing findings, we confirmed the reproducibility and strength of our models.

9. Conclusion

In conclusion, each methodology shows an exciting insight into the dataset. This project uses the Insurance dataset from Kaggle. With only seven columns and 1339 rows of data, we have leveraged all of it to execute different methodologies such as linear regression, logistic regression, k-NN model, and decision tree. In addition, we also plotted some basic histograms and boxplots to sketch our initial thoughts.

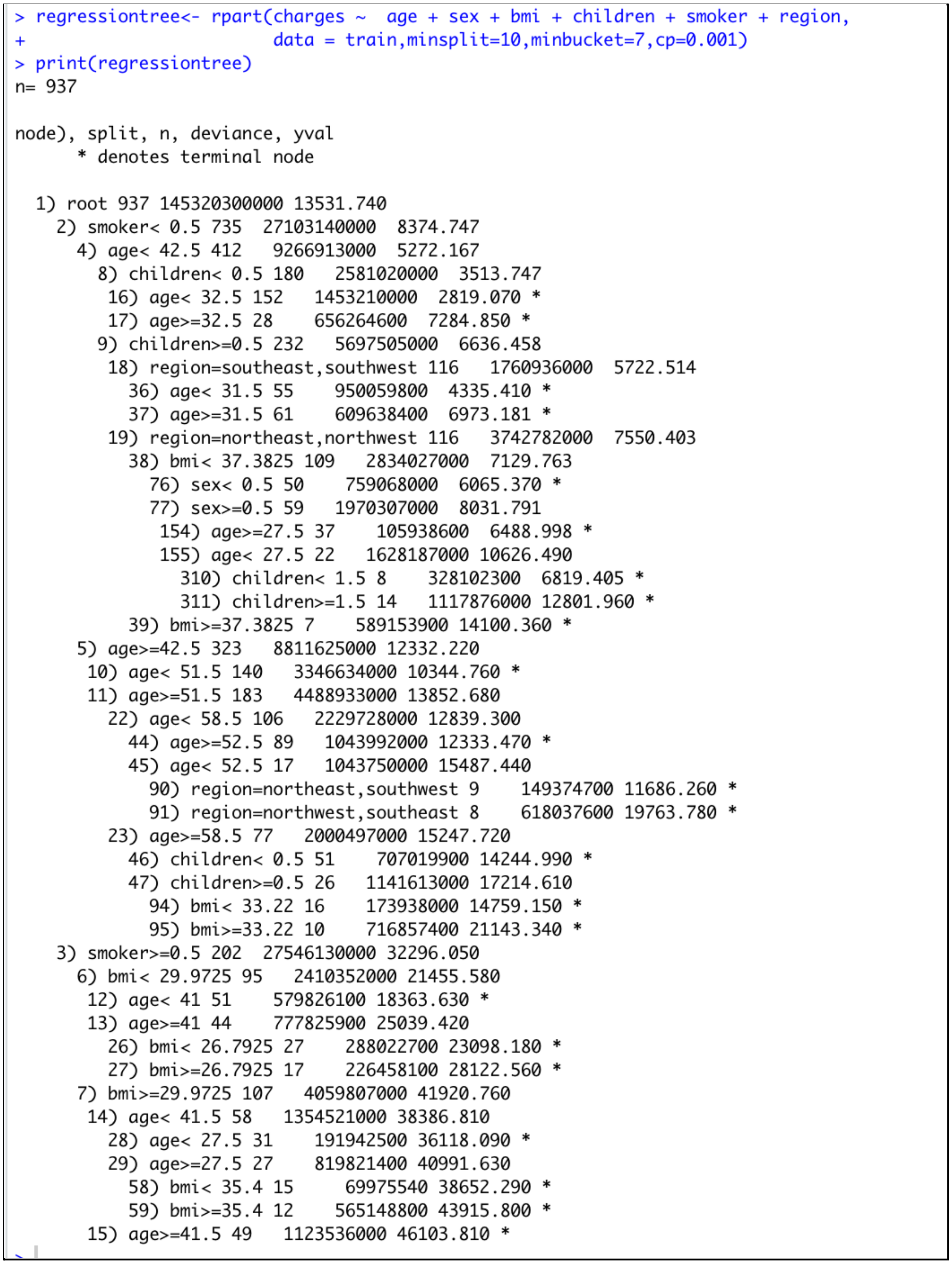
At first, the group thought age, BMI, and the number of children would impact insurance charges by looking at the histograms and boxplots. After running advanced methodologies, our final thought has been made. Logistic regression shows the best accuracy than k-NN, and logistic regression also contains reasonable estimations of RMSE and AE.

The decision tree shows a detailed view of which factor has the highest significance with a number. This helps us to predict insurance charges for different types of customers. Insurance companies could use this to deliver insurance charges faster that can satisfy customers' demands.

These results above can be helpful for companies to take advantage of in real-life situations. Nevertheless, there are more advanced analytic tools to make better predictions. Therefore, companies, especially data scientists, should invest time to find out which methodology works best for their business and customers' goals.

10. Appendix

Appendix 1: Decision Tree Node



```
> regressiontree<- rpart(charges ~  age + sex + bmi + children + smoker + region,
+                        data = train,minsplit=10,minbucket=7,cp=0.001)
> print(regressiontree)
n= 937

node), split, n, deviance, yval
      * denotes terminal node

  1) root 937 145320300000 13531.740
    2) smoker< 0.5 735  27103140000  8374.747
      4) age< 42.5 412   9266913000  5272.167
        8) children< 0.5 180   2581020000  3513.747
         16) age< 32.5 152   1453210000  2819.070 *
         17) age>=32.5 28    656264600  7284.850 *
        9) children>=0.5 232   5697505000  6636.458
         18) region=southeast,southwest 116   1760936000  5722.514
           36) age< 31.5 55    950059800  4335.410 *
           37) age>=31.5 61    609638400  6973.181 *
         19) region=northeast,northwest 116   3742782000  7550.403
           38) bmi< 37.3825 109   2834027000  7129.763
             76) sex< 0.5 50    759068000  6065.370 *
             77) sex>=0.5 59   1970307000  8031.791
              154) age>=27.5 37    105938600  6488.998 *
              155) age< 27.5 22   1628187000 10626.490
                310) children< 1.5 8    328102300  6819.405 *
                311) children>=1.5 14   1117876000 12801.960 *
           39) bmi>=37.3825 7    589153900 14100.360 *
      5) age>=42.5 323   8811625000 12332.220
       10) age< 51.5 140   3346634000 10344.760 *
       11) age>=51.5 183   4488933000 13852.680
         22) age< 58.5 106   2229728000 12839.300
           44) age>=52.5 89   1043992000 12333.470 *
           45) age< 52.5 17   1043750000 15487.440
             90) region=northeast,southwest 9    149374700 11686.260 *
             91) region=northwest,southeast 8    618037600 19763.780 *
         23) age>=58.5 77   2000497000 15247.720
           46) children< 0.5 51    707019900 14244.990 *
           47) children>=0.5 26   1141613000 17214.610
             94) bmi< 33.22 16    173938000 14759.150 *
             95) bmi>=33.22 10    716857400 21143.340 *
    3) smoker>=0.5 202  27546130000 32296.050
      6) bmi< 29.9725 95   2410352000 21455.580
       12) age< 41 51    579826100 18363.630 *
       13) age>=41 44    777825900 25039.420
         26) bmi< 26.7925 27    288022700 23098.180 *
         27) bmi>=26.7925 17    226458100 28122.560 *
      7) bmi>=29.9725 107   4059807000 41920.760
       14) age< 41.5 58   1354521000 38386.810
         28) age< 27.5 31    191942500 36118.090 *
         29) age>=27.5 27    819821400 40991.630
           58) bmi< 35.4 15     69975540 38652.290 *
           59) bmi>=35.4 12    565148800 43915.800 *
       15) age>=41.5 49   1123536000 46103.810 *
>
```