FinalProject

2023-04-11

Attributes of Billed Health Insurance Costs

Logistic Regression Classifier and kNN Model

Report by Abena Boateng and Sedem Kakrada

Introduction

The dataset is known as US Health Insurance Dataset retrieved from Kaggle at: https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset (https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset)

Within this dataset the insurance charge is listed and given with other attributes Age, Sex, BMI, Number of Children, Smoker and Region with a total of 1338 patients as observations.

Through the creation of classification and prediction models we expected there to be attributes of more significance that directly effect the effectiveness of the models produced. The biggest predictors are expected to be Age, BMI, Children and Smoker Status each being positively related. These stronger correlations are expected to aid in the best classification models being produced from this exploration.

The Research Questions to be addressed are:

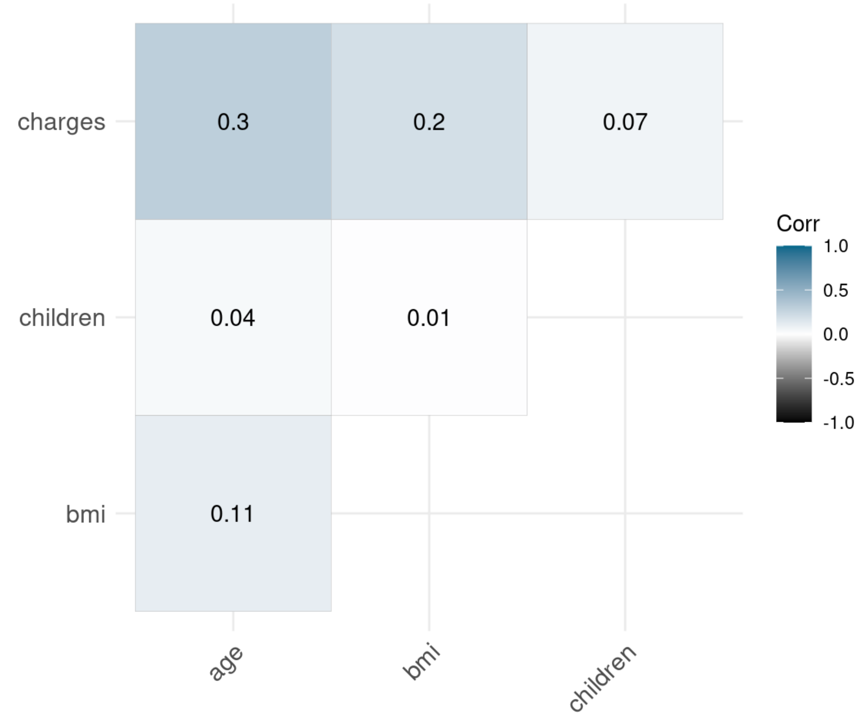
Are the age and body mass index of a primary beneficiary member significant predictors of the individual medical costs billed by health insurance companies?

Is smoking status, regional area, and number of children of the beneficiary member on an insurance plan effective classifiers for the medical costs billed by the health insurance companies?

Exploratory Data Analysis

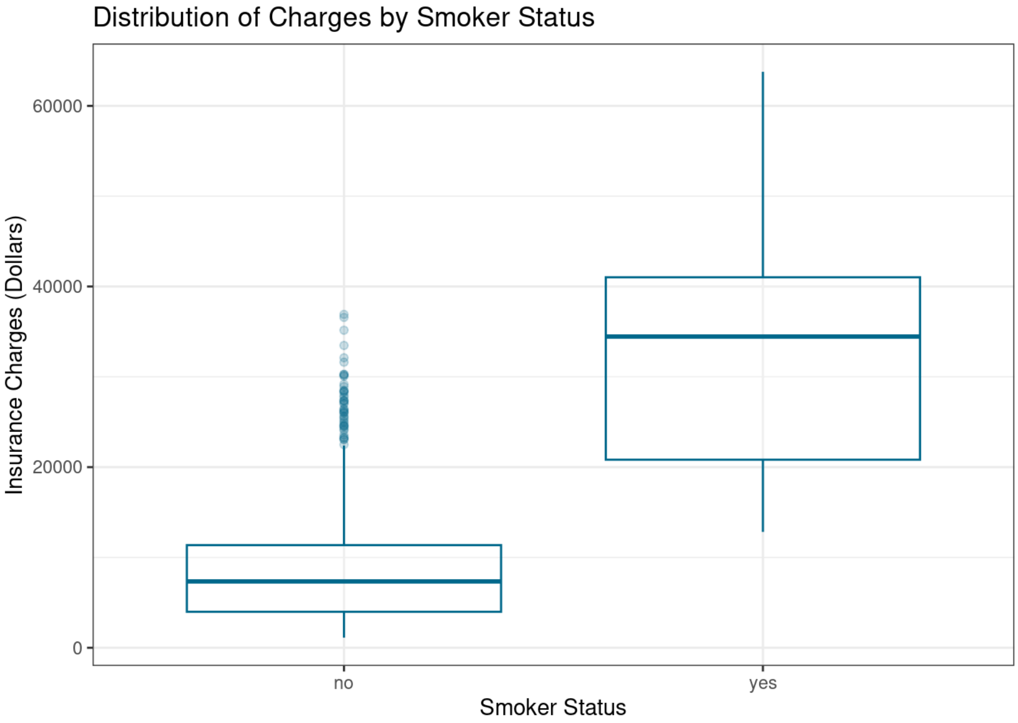
Creating a Correlation Matrix

```
#correlation of numeric variables in insurance dataset
ggcorrplot(cor(insurance %>%
                 select(-c(sex, smoker, region))),
           lab = TRUE, colors = c('black', 'white', 'deepskyblue4'),
           type = 'upper')
```

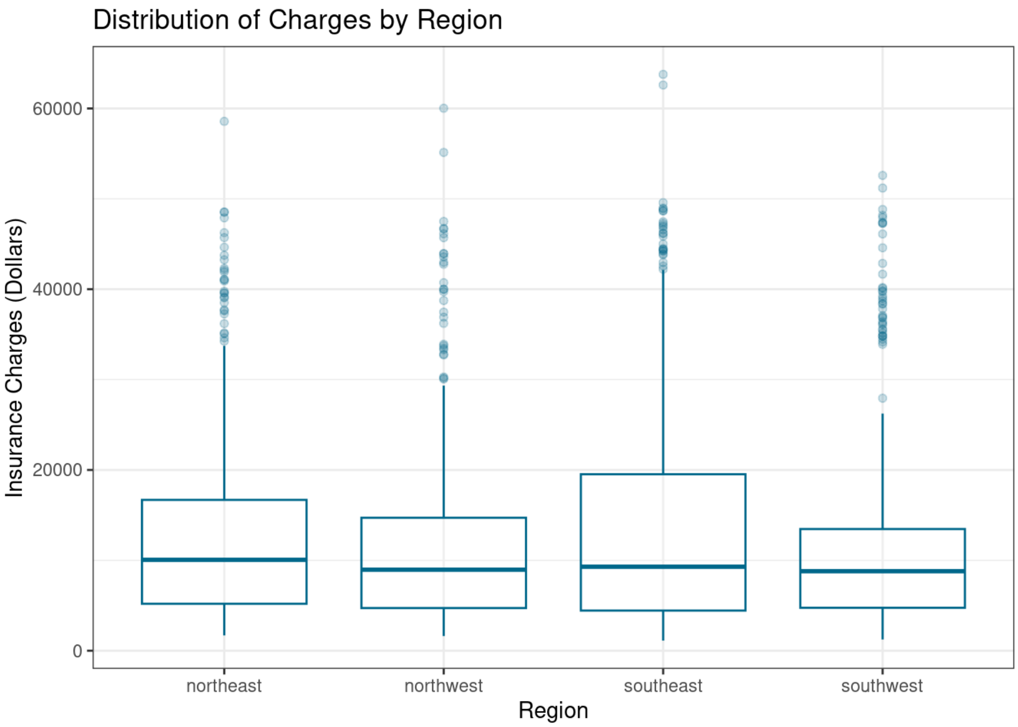


Investigation of Variables Through Visualization

```
#investigation of smoking effect on total insurance charges
insurance %>%
  ggplot(aes(x = smoker, y = charges)) +
  geom_boxplot(color="deepskyblue4", fill="white", alpha=0.2) +
  theme_bw() +
  labs(x = 'Smoker Status', y = 'Insurance Charges (Dollars)', title = 'Distribution of Charges by Smoker Statu
s')
```

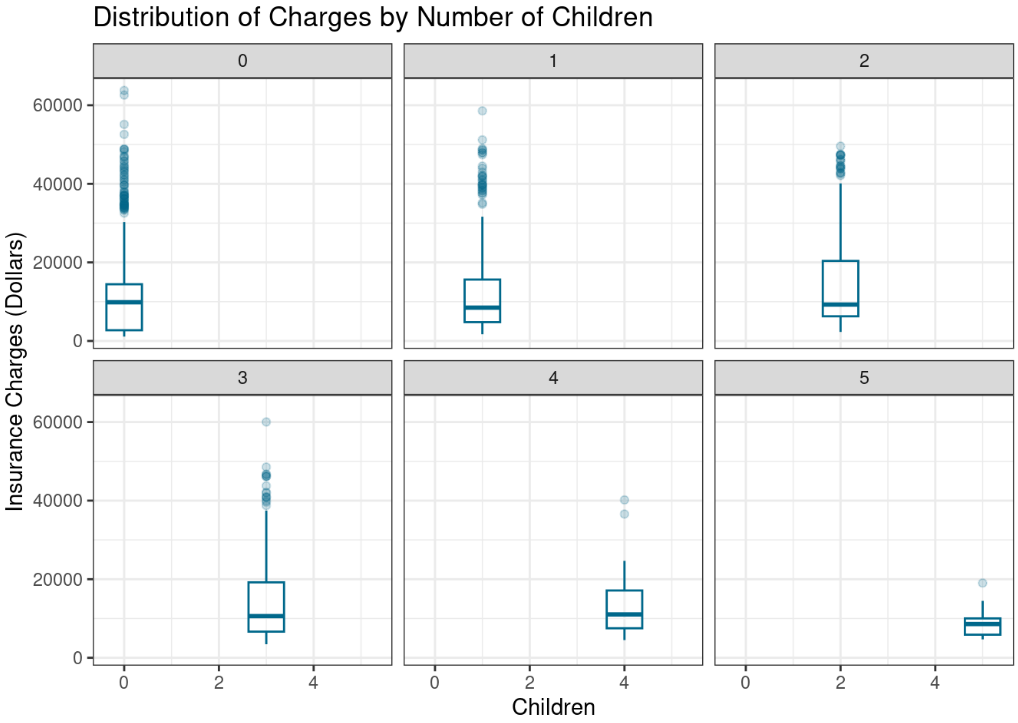


```
#investigation of region effect on total insurance charge
insurance %>%
  ggplot(aes(x = region, y = charges)) +
  geom_boxplot(color="deepskyblue4", fill="white", alpha=0.2)+
  theme_bw() +
  labs(x = 'Region', y = 'Insurance Charges (Dollars)', title = 'Distribution of Charges by Region')
```



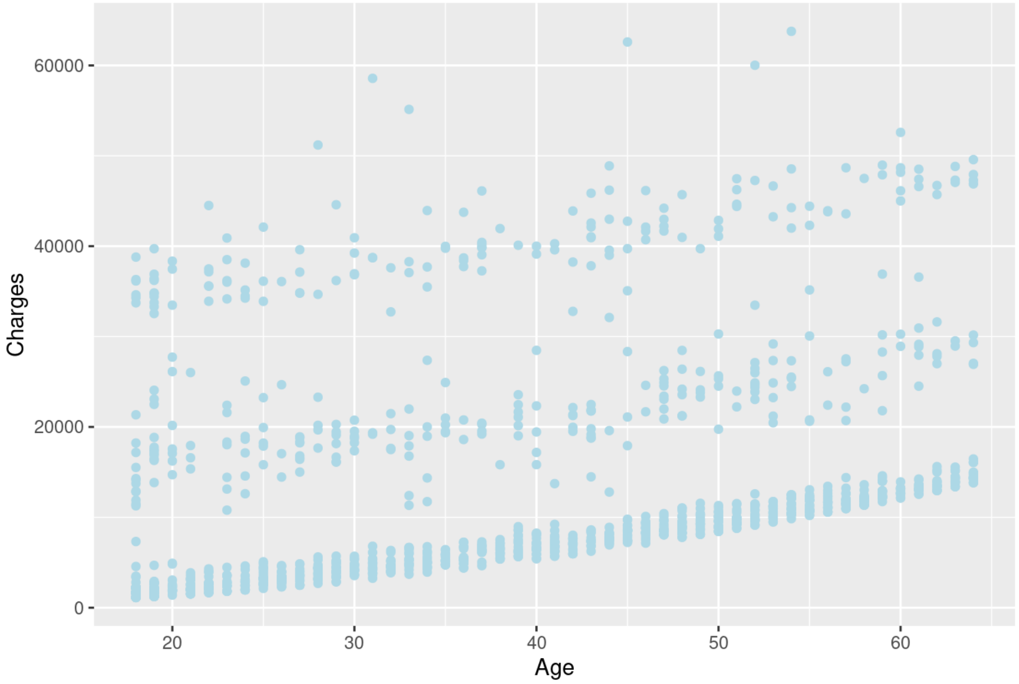
```
#investigation of children effect on total insurance charge
insurance %>%
  ggplot(aes(x = children, y = charges)) +
  geom_boxplot(color="deepskyblue4", fill="white", alpha=0.2)+
  theme_bw() +
  labs(x = 'Children', y = 'Insurance Charges (Dollars)', title = 'Distribution of Charges by Number of Children') +
  facet_wrap(~children)
```

```
## Warning: Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?
```

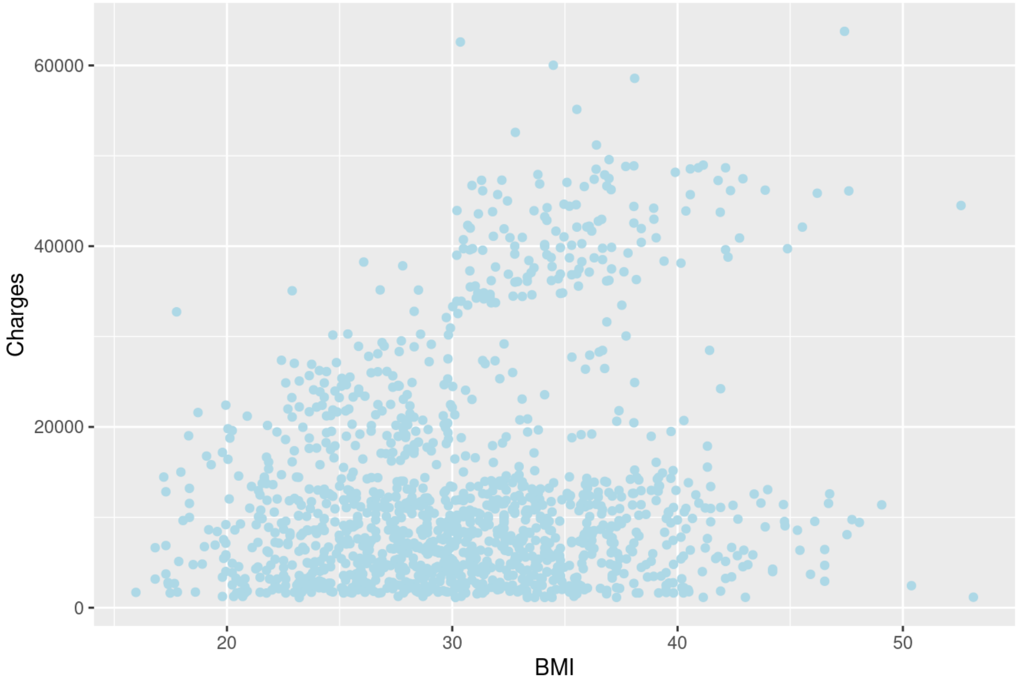


Through a box plot visualization of distribution, the effect of smoking appears to be a significant factor is the predicted insurance charges. This was to be expected as those with longer history of smoking have more health related risk and conditions that contribute to more frequent health complication resulting in higher insurance billings.The distribution of insurance charges remain in fairly the same ranges despite the region of the country in which the patient was billed. The region with the widest distribution is the southeast with it still being right skewed. After accounting for cost of living among the respective regions of the country, it is reasonable that the difference among the regions do not have significant impact. Looking at the distribution of insurance charges and the number of children by the beneficiary, there is a noticeable trend where as the number of children increase, the range of billed insurance charges decrease. Despite this trend the median of each children number group remain fairly the same. This pattern makes sense as those with larger families often have better coverage deals and plans provided by the insurance to have the burden of cost be less while still widening the customer space.

```
#exploring relationship of charge based on age
insurance %>%
  ggplot(aes(x = age, y = charges)) +
  geom_point(color = "light blue") +
  labs(title = "Relationship of charge based on Age Fig 1.1",
       x = "Age",
       y = "Charges")
```


Relationship of charge based on Age Fig 1.1

```
#exploring relationship of charge based on bmi
insurance %>%
  ggplot(aes(x = bmi, y = charges)) +
  geom_point(color = "light blue") +
  labs(title = "Relationship of charge based on Age Fig 1.2",
       x = "BMI",
       y = "Charges")
```


Relationship of charge based on Age Fig 1.2

Based upon our second research question’s aim to explore the relationship between Age, BMI, and their influence on insurance price charges, it was vital to create a visualization. As a result of two numeric variables, a scatter plot was generated to explore the relationships between Age, BMI, and Charges. Principally, the first scatter plot explores the relationship between Age and Charges, showcasing that there is somewhat of a positive relationship between these two variables. In addition to this, the relationship between BMI and charges appears to be not as apparent.

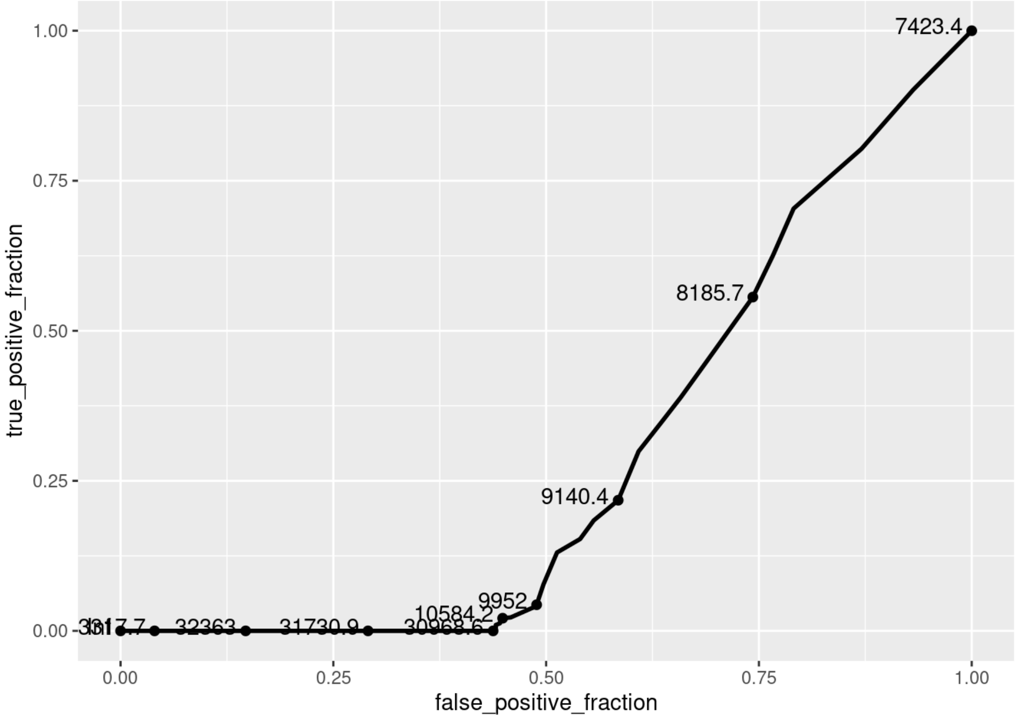
Prediction and Cross-Validation

Classification Model

```
#log regression model
insurance_class <- insurance %>% mutate(charges_class = ifelse(charges < 10000, 'low', 'high'))
insurance_log <- glm(charges ~ children + region + smoker, data = insurance_class)
summary(insurance_log)
```

```
## 
## Call:
## glm(formula = charges ~ children + region + smoker, data = insurance_class)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -19856   -4978   -1144    3919   32040  
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       7876.0      460.9  17.087  < 2e-16 ***
## children           632.2      168.9   3.744 0.000189 ***
## regionnorthwest   -385.9      584.3  -0.660 0.509064    
## regionsoutheast    309.6      568.4   0.545 0.586031    
## regionsouthwest   -452.7      584.3  -0.775 0.438624    
## smokeryes        23545.2      505.2  46.602  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 55309060)
## 
##     Null deviance: 1.9607e+11  on 1337  degrees of freedom
## Residual deviance: 7.3672e+10  on 1332  degrees of freedom
## AIC: 27660
## 
## Number of Fisher Scoring iterations: 2
```

```
#ROC curve for log regression classification
log_ROC <- insurance_class %>%
  mutate(predictions = predict(insurance_log, type = "response"), class_predict = ifelse(predictions < 10000, 'lo
w', 'high')) %>%
  ggplot() +
  geom_roc(aes(d = charges_class, m = predictions), n.cuts = 10) +
  labs(xlab = 'ROC Curve for Insurance Dataset Log Regression Model')
log_ROC
```



```
#AUC of log regression model
calc_auc(log_ROC)$AUC
```

```
## [1] 0.2863463
```

The logistic regression model was used to classify the charges build into two categories low or high based on the 10,000 threshold in which it an AUC of ~0.28 was produced. Although the best approximate choice for threshold was utilized, this regression classification model is not significantly effective at utilizing the logistic equation to predict the outcome charges and correctly classifying them for new observations.

```
#logistic regression model
summary(insurance_log)
```

```
## 
## Call:
## glm(formula = charges ~ children + region + smoker, data = insurance_class)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -19856   -4978   -1144    3919   32040  
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       7876.0      460.9  17.087  < 2e-16 ***
## children           632.2      168.9   3.744 0.000189 ***
## regionnorthwest   -385.9      584.3  -0.660 0.509064    
## regionsoutheast    309.6      568.4   0.545 0.586031    
## regionsouthwest   -452.7      584.3  -0.775 0.438624    
## smokeryes        23545.2      505.2  46.602  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 55309060)
## 
##     Null deviance: 1.9607e+11  on 1337  degrees of freedom
## Residual deviance: 7.3672e+10  on 1332  degrees of freedom
## AIC: 27660
## 
## Number of Fisher Scoring iterations: 2
```

```
#RMSE
insurance_RMSE <- insurance_class %>%
  mutate(predictions = predict(insurance_log, type = "response"))
sqrt(mean((insurance_RMSE$charges - insurance_RMSE$predictions)^2))
```

```
## [1] 7420.312
```

The equation created by the logisitic regression model of the variables smoker status, region, and number of children to predict the insurance cost is:

$$\ln \frac{\hat{p}}{1-\hat{p}} = 7876.0 + 632.2 * children - 385.9 * regionnorthwest + 309.6 * regionsoutheast - 452.7 * regionsouthwest + 23545.2 * smokery$$

After utilizing the equation to make these predictions, the root mean square estimate is used to average error by the model based on residuals which was found to be 7420.312 dollars billed by health insurance companies.

Investigating with kNN method

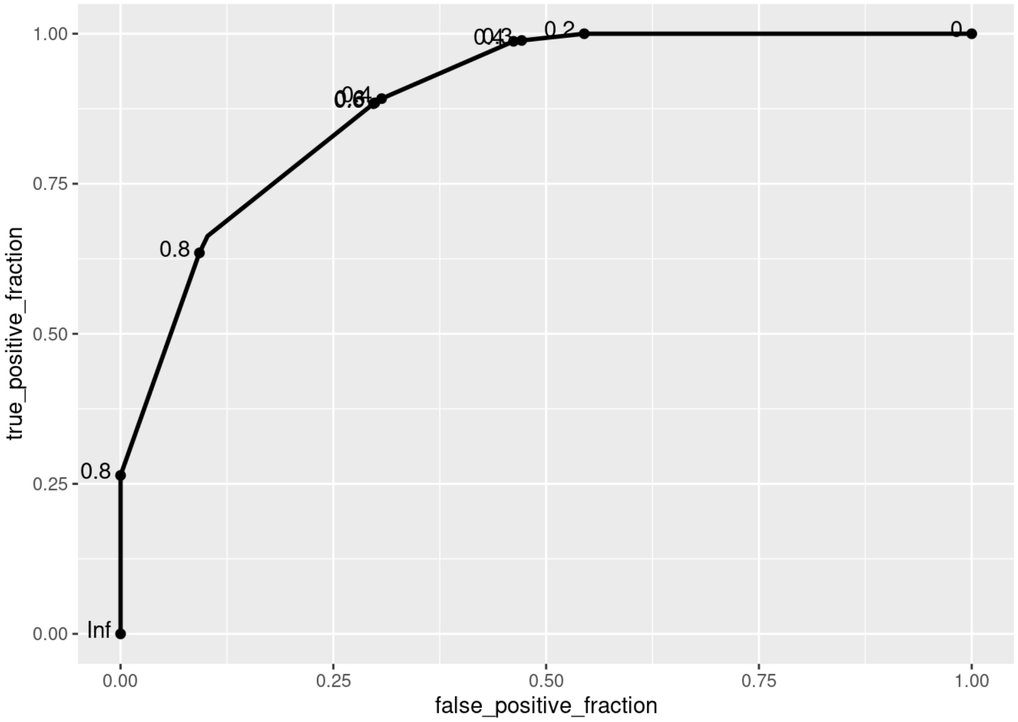
```
#Consider the kNN Classifier
insurance_kNN <- knn3(charges_class ~ bmi + age,
                data = insurance_class,
                k = 5)
predict(insurance_kNN, insurance_class) %>% as.data.frame %>% head
```

```
##   high low
## 1  0.6 0.4
## 2  0.4 0.6
## 3  0.2 0.8
## 4  0.6 0.4
## 5  0.4 0.6
## 6  0.2 0.8
```

```
insurance %>%
  mutate(predictions = predict(insurance_kNN, insurance_class)[,2], # keep column 2
         predicted = ifelse(predictions > 1000, "high","low"))
```

```
## # A tibble: 1,338 × 9
##      age sex      bmi children smoker region    charges predictions predicted
##    <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>       <dbl> <chr>
##  1    19 female  27.9        0 yes    southwest  16885.         0.4 low
##  2    18 male    33.8        1 no     southeast   1726.         0.6 low
##  3    28 male    33          3 no     southeast   4449.         0.8 low
##  4    33 male    22.7        0 no     northwest  21984.         0.4 low
##  5    32 male    28.9        0 no     northwest   3867.         0.6 low
##  6    31 female  25.7        0 no     southeast   3757.         0.8 low
##  7    46 female  33.4        1 no     southeast   8241.         0.8 low
##  8    37 female  27.7        3 no     northwest   7282.         1   low
##  9    37 male    29.8        2 no     northeast   6406.         0.6 low
## 10    60 female  25.8        0 no     northwest  28923.         0   low
## # … with 1,328 more rows
```

```
ROC <- ggplot(insurance_class) +
  geom_roc(aes(d = charges_class, m = predict(insurance_kNN, insurance_class)[,2]), n.cuts = 10)
ROC
```



```
calc_auc(ROC)
```

```
##   PANEL group      AUC
## 1     1    -1 0.890455
```

Using our suite of classification tools, I decided to perform a K- Nearest Neighbors algorithm to explore our data. This algorithm bases the class of a data point on the majority voting principle. If k we make equal to 5, the classes of 5 closest points are checked and the prediction is done according to the majority class. After setting up our classifier to create predictions based on two categories, high and low based on the charge amount, we then visualized the prediction model using a ROC curve. By means of calculating, we got an AUC of ~0.89 showcasing that the kNN algorithm is an adequate method of classification for this dataset.

Based on the AUC values of each model, the kNN model utilizing age and BMI is significantly more effective than the logistic regression model utilizing smoking status, region, and number of children.

Cross-Validation

Logistic Cross-Validation

```
# Choose number of folds
k = 10

# Randomly order rows in the dataset
data <- insurance_class[sample(nrow(insurance_class)), ]

# Create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)

perf_k_insurance <- NULL

# Use a for loop to get diagnostics for each test set
for(i in 1:k){
  # Create train and test sets
  train_not_i <- data[folds != i, ] # all observations except in fold i
  test_i <- data[folds == i, ]  # observations in fold i

  # Train model on train set (all but fold i)
  insurance_log <- glm(charges ~ children + region + smoker, data = train_not_i %>%
                         mutate(charges_class = ifelse(charges < 10000, 'low', 'high')))

  # Test model on test set (fold i)
  predict_i <- data.frame(
    predictions = predict(insurance_log, newdata = test_i, type = "response"),
    charges_class = test_i$charges_class)

  # Consider the ROC curve for the test dataset
  ROC <- ggplot(predict_i) +
    geom_roc(aes(d = charges_class, m = predictions))

  # Get diagnostics for fold i (AUC)
  perf_k_insurance[i] <- calc_auc(ROC)$AUC
}
```

```
mean(perf_k_insurance)
```

```
## [1] 0.2879215
```

Although the model it self is not significantly effective at the objective, it is very reproducible. Through the cross validation process an AUC of approximately the same value was reproduced on the test data, limiting the idea of over fitting issues with the model.

kNN Cross-Validation

```
# Choose number of folds
k = 10

# Randomly order rows in the dataset
data <- insurance_class[sample(nrow(insurance_class)),]

# Create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)


perf_k_insurance <- NULL

# Use a for loop to get diagnostics for each test set
for(i in 1:k){
  # Create train and test sets
  train <- data[folds != i, ] # all observations except in fold i
  test <- data[folds == i, ]  # observations in fold i

  # Train model on train set (all but fold i)
  insurance_kNN <- knn3(charges_class ~ bmi + age,
              data = train,
              k = 5)

  # Test model on test set (fold i)
  df <- data.frame(
    predictions = predict(insurance_kNN, test)[,2],
    charges_class = test$charges_class)

  # Consider the ROC curve for the test dataset
  ROC <- ggplot(df) +
    geom_roc(aes(d = charges_class, m = predictions))

  # Get diagnostics for fold i (AUC)
  perf_k_insurance[i] <- calc_auc(ROC)$AUC
}
```

```
# Average performance
mean(perf_k_insurance)
```

```
## [1] 0.7602736
```

In addition to the kNN Algorithm a cross validation test was performed to showcase how well our model would work on new data with some slight over fitting issues. As a result, our model exhibits an average performance of 0.76 signifying it’s ability to work well on new data, but work best on the original train data. Overall our model is pretty accurate and should be used to explore how different variables would perform on this data set.

Dimension Reduction

```
insurance_class_scaled <- insurance_class %>%
  select_if(is.numeric)%>%
  scale%>%
  as.data.frame()

head(insurance_class_scaled)
```

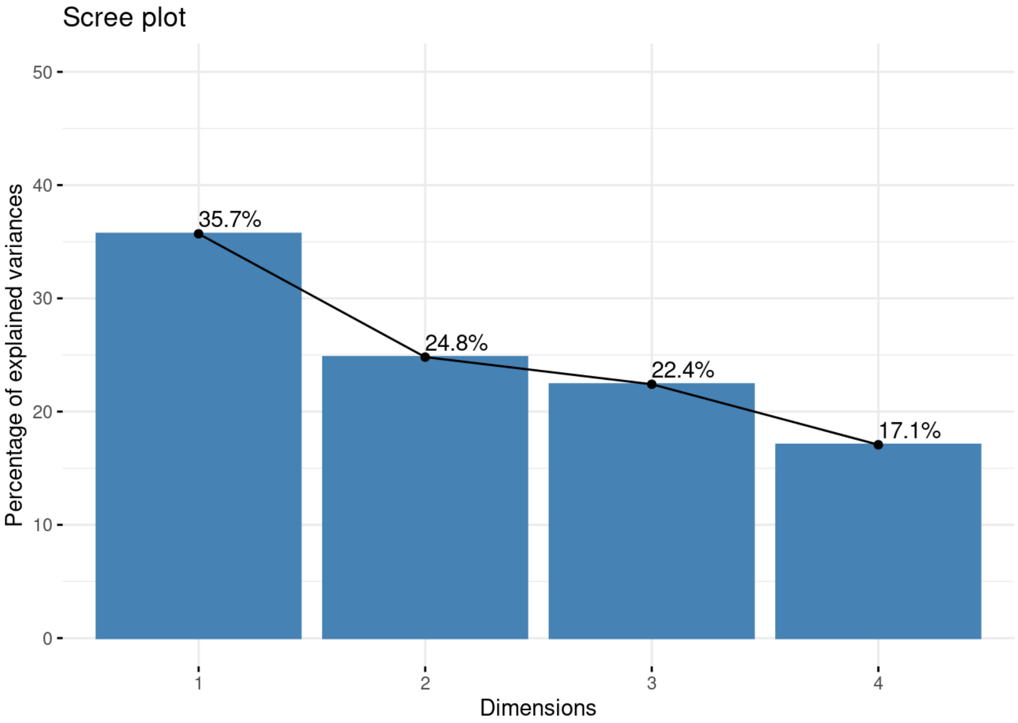
```
##          age        bmi    children    charges
## 1 -1.4382265 -0.4531506 -0.90827406  0.2984722
## 2 -1.5094011  0.5094306 -0.07873775 -0.9533327
## 3 -0.7976553  0.3831636  1.58033487 -0.7284023
## 4 -0.4417824 -1.3050431 -0.90827406  0.7195739
## 5 -0.5129570 -0.2924471 -0.90827406 -0.7765118
## 6 -0.5841316 -0.8073542 -0.90827406 -0.7856145
```

```
pca <- insurance_class_scaled %>%
  prcomp

names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

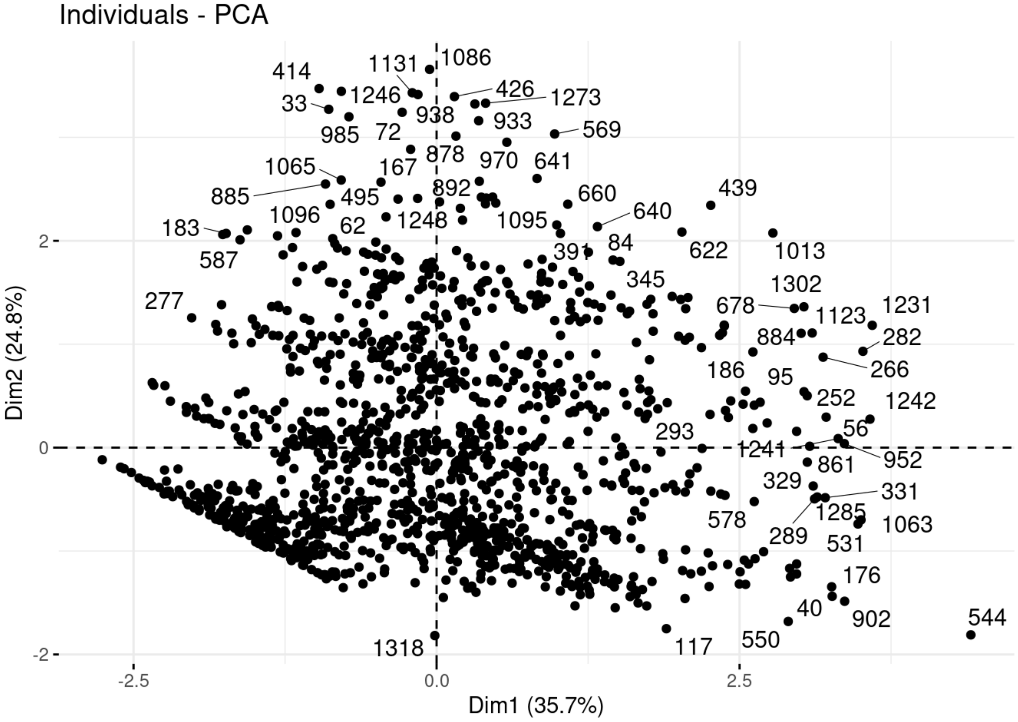
```
pca$x %>% as.data.frame
```

```
fviz_eig(pca, addlabels = TRUE, ylim = c(0, 50))
```



```
fviz_pca_ind(pca,
             repel = TRUE)
```

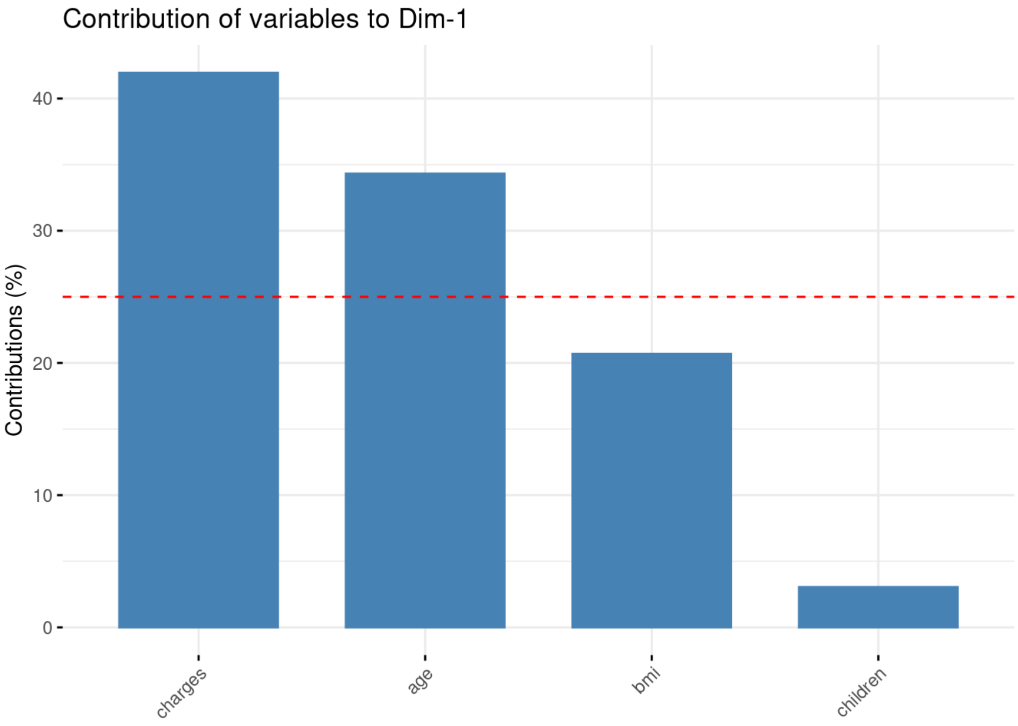
```
## Warning: ggrepel: 1273 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



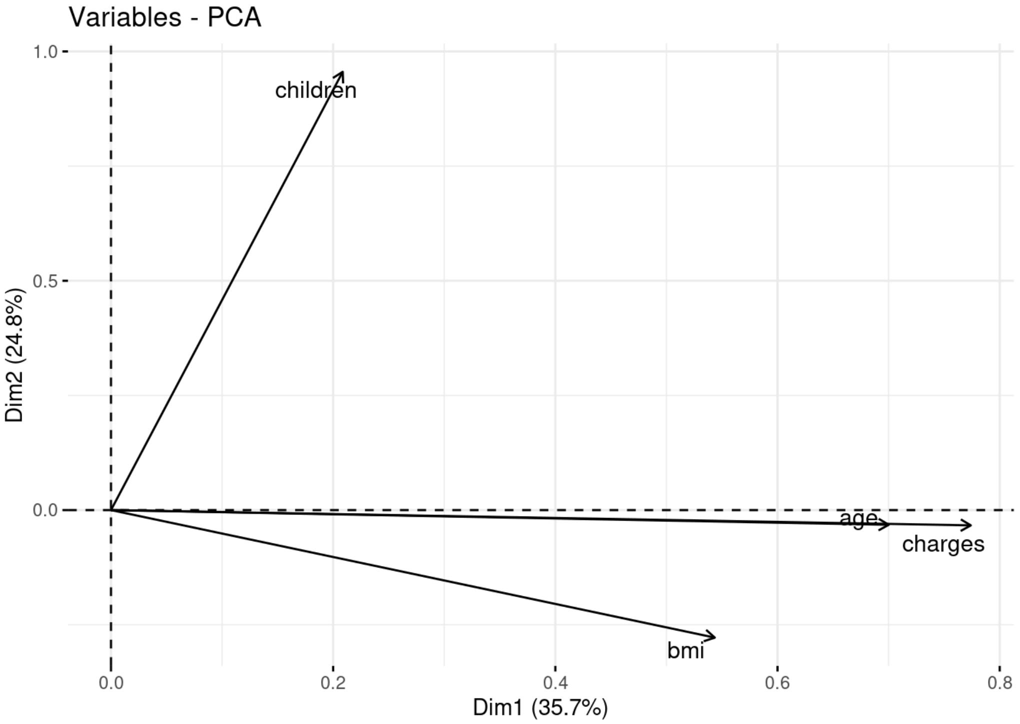
```
get_pca_var(pca)$coord %>% as.data.frame
```

```
##                Dim.1       Dim.2      Dim.3      Dim.4
## age      0.7000654 -0.03202662 -0.5061502  0.5026874
## bmi      0.5435330 -0.27804918  0.7632993  0.2112693
## children 0.2085990  0.95553550  0.2005366  0.0567753
## charges  0.7739741 -0.03330106 -0.1322677 -0.6183529
```

```
fviz_contrib(pca, choice = "var", axes = 1, top = 4)
```



```
fviz_pca_var(pca, col.var = "black",
             repel = TRUE)
```



A PCA, further known as a Principal Component Analysis, functions by preparing our insurance data through scaling, performing a PCA, making a scree plot, and lastly considering the PC Score. In relation to our dataset, our scree plots showcase 4 main principle components in which the top 3 showcase an 82.9% majority of our variation. Along with the scree plot, a visualization was created in order to visualize which 3 variables mainly contribute to the dimensions showcased in our plot. As a result, it is evident that the charges and age contribute mostly to our dimensions whereas BMI and Children do not. To further explain scoring high on the first two PCs showcases not only the variation but also the possible significance of each associated variable within our data set.

Clustering

KMeans Clustering

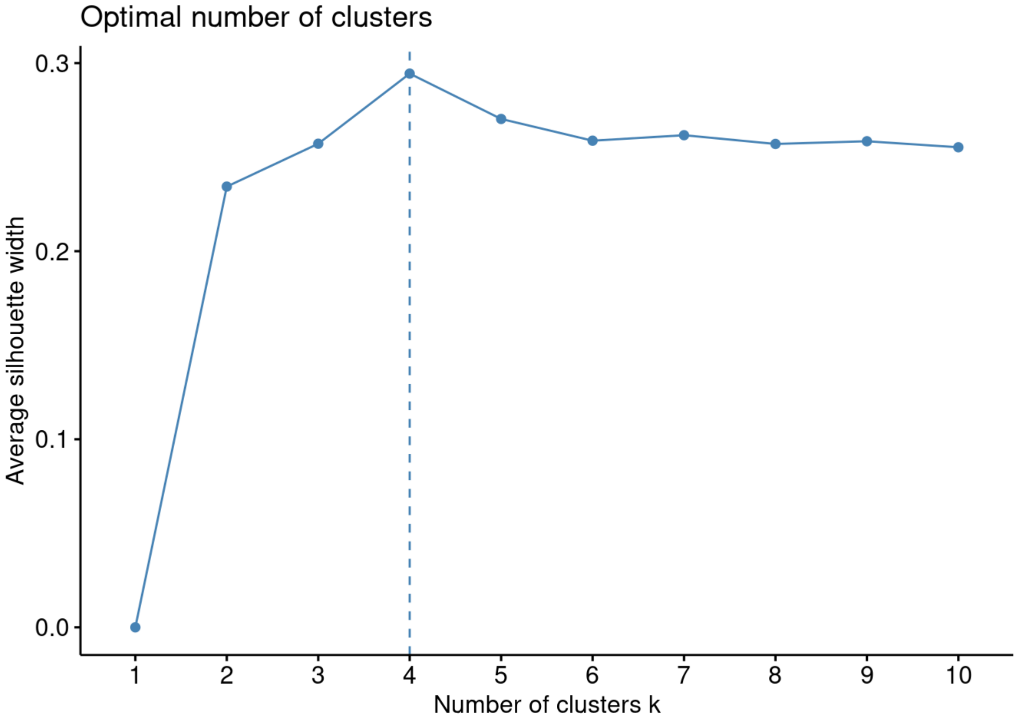
```
# Keep two variables and scale them
insurance_scaled <- insurance_class %>%
  select_if(is.numeric) %>%
  scale

# Use the function kmeans() to find clusters
kmeans_results <- insurance_scaled %>%
  kmeans(centers = 4) # centers sets the number of clusters to find

kmeans_results$centers
```

```
##           age         bmi    children     charges
## 1 -0.98400020 -0.24694467 -0.56576862 -0.65012653
## 2  0.04738105  0.76079694  0.04192208  2.23268973
## 3  0.05211979 -0.11479074  1.31900696 -0.21146267
## 4  0.94716854  0.04327431 -0.55450122 -0.05594291
```

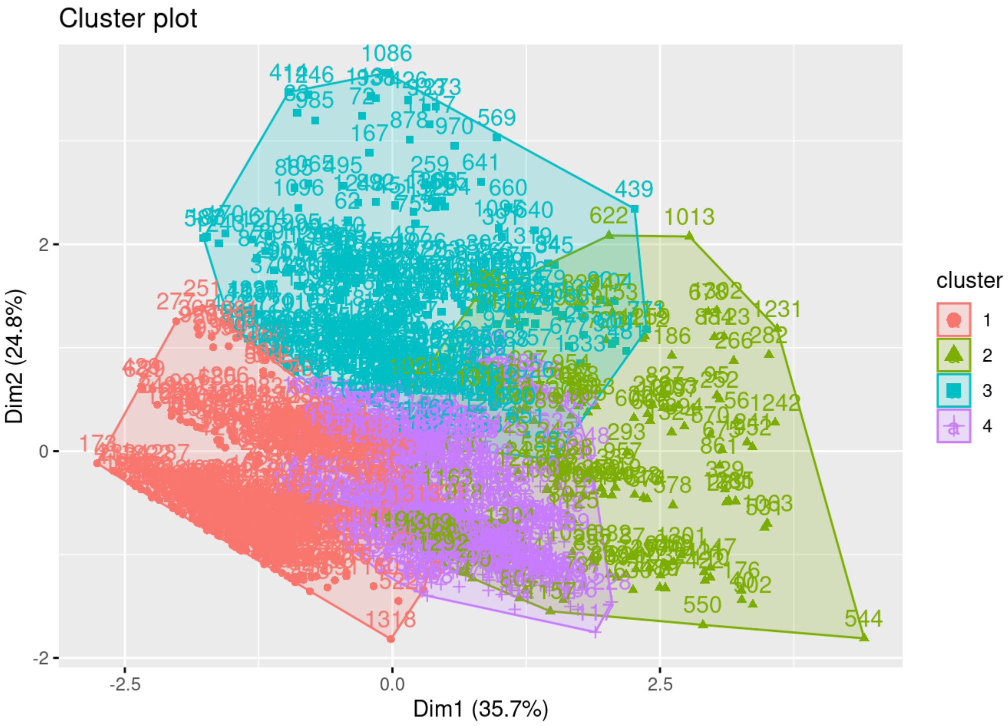
```
#Save cluster as its own column
insurance %>%
  mutate(cluster = as.factor(kmeans_results$cluster)) %>%
  head
```

```
## # A tibble: 6 × 8
##     age sex      bmi children smoker region    charges cluster
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl> <fct>  
## 1    19 female  27.9        0 yes    southwest  16885. 1      
## 2    18 male    33.8        1 no     southeast   1726. 1      
## 3    28 male    33          3 no     southeast   4449. 3      
## 4    33 male    22.7        0 no     northwest  21984. 1      
## 5    32 male    28.9        0 no     northwest   3867. 1      
## 6    31 female  25.7        0 no     southeast   3757. 1
```

```
fviz_nbclust(insurance_scaled, kmeans, method = "silhouette")
```



```
fviz_cluster(kmeans_results, data = insurance_class_scaled)
```



```
insurance_class_scaled %>%
  select(age, bmi) %>%
  mutate(cluster = as.factor(kmeans_results$cluster)) %>%
  group_by(cluster) %>%
  summarize_if(is.numeric, mean, na.rm = T)
```

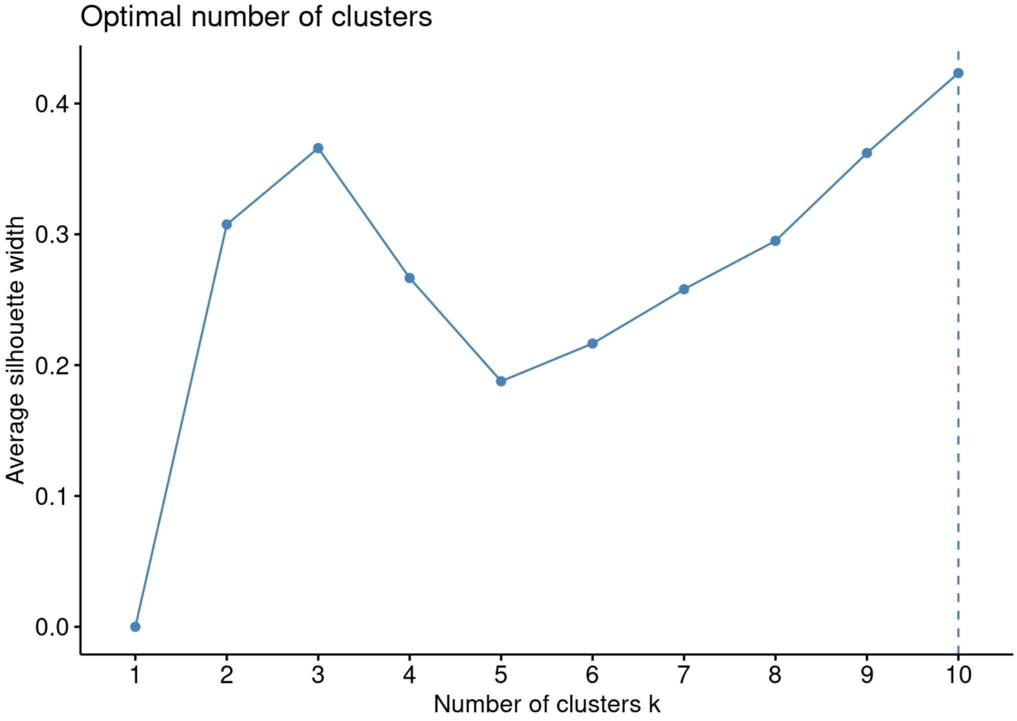
```
## # A tibble: 4 × 3
##   cluster     age     bmi
##   <fct>     <dbl>   <dbl>
## 1 1       -0.984  -0.247
## 2 2        0.0474  0.761
## 3 3        0.0521 -0.115
## 4 4        0.947   0.0433
```

For this dataset a k-means clustering algorithm was performed on all numeric variables within our insurance dataset. When applying the algorithm we find that we have 4 clusters which has sizes of 419,165,346, and 408. First and foremost, our first cluster has a center of -0.98 for age, -0.24 for BMI, and -0.65 for charges. We observe that cluster one typically consists more of negative means. On the other hand, cluster 2 exhibits a a center of 0.04 for age, 0.76 for BMI, and 2.2 for charges. This further explains how cluster 2 individuals who are further along in age and BMI typically showcase higher insurance charges. In terms of cluster 3 a center of 0.052 for age, -0.11 for BMI, and -0.21 for charges. Individuals belonging to cluster 3 are also farther along in age but do not seem to experience higher charges as a result. Last but not least, cluster 4 showcases a mean center of 0.94 for age , -0.02 for charges and a -0.11 for BMI which could point to this cluster representing individuals with a typically lower BMI , who are father along in age experiencing lower charges.. Overall we witness that the average silhouette width seems to point towards having 4 clusters to maximize our average silhouette width.

Gower PAM Clustering

```
#insurance dataset w/o numeric variables
insurance_cat <- insurance %>%
  mutate_if(is.character, as.factor) %>%
  na.omit

#saving insurance in gower matrix
insurance_cat %>%
  daisy(metric = "gower") %>%
  as.matrix -> insurance_gower
```

```
fviz_nbclust(insurance_gower, pam, method = "silhouette")
```



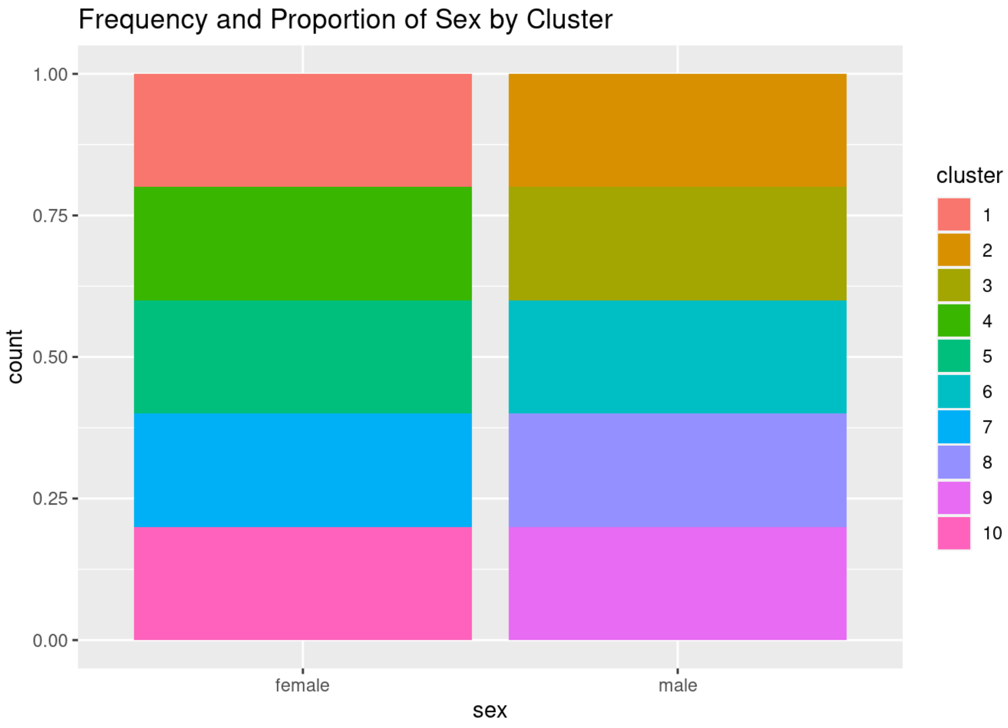
```
insurance_gpam <- pam(insurance_gower, k = 10, diss = TRUE)
```

```
#mean of numerical variables by cluster for gower dissimilarity
insurance %>%
  mutate(cluster = as.factor(insurance_gpam$clustering)) %>%
  group_by(cluster) %>%
  summarize_if(is.numeric, mean, na.rm = T)
```

```
## # A tibble: 10 × 5
##    cluster   age   bmi children charges
##    <fct>   <dbl> <dbl>    <dbl>   <dbl>
##  1 1        39.5  29.7    1.13    8874.
##  2 2        38.3  34.1    1.05    7609.
##  3 3        38.4  28.7    1.10    9707.
##  4 4        39.1  32.8    1.08    8440.
##  5 5        39.2  29.1    1.16    9916.
##  6 6        38.9  28.5    1.10    9952.
##  7 7        41.1  32.9    0.919  36834.
##  8 8        39.8  30.6    1.18    9026.
##  9 9        39.3  33.6    1.15   38157.
## 10 10       39.4  28.9    1      10757.
```

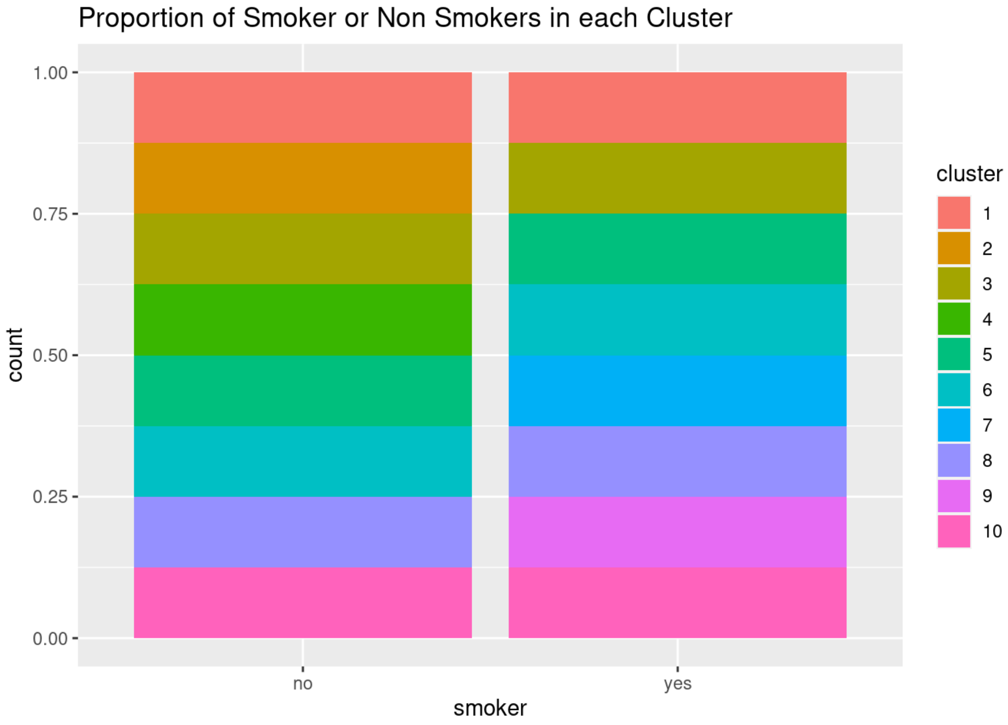
```
#frequency and proportion of sex by cluster
insurance %>%
  mutate(cluster = as.factor(insurance_gpam$clustering)) %>%
  group_by(cluster, sex) %>%
  summarize(freq = n()) %>%
  ggplot(aes( x = sex, fill = cluster)) +
  geom_bar(position = 'fill') +
  labs(title = 'Frequency and Proportion of Sex by Cluster')
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```



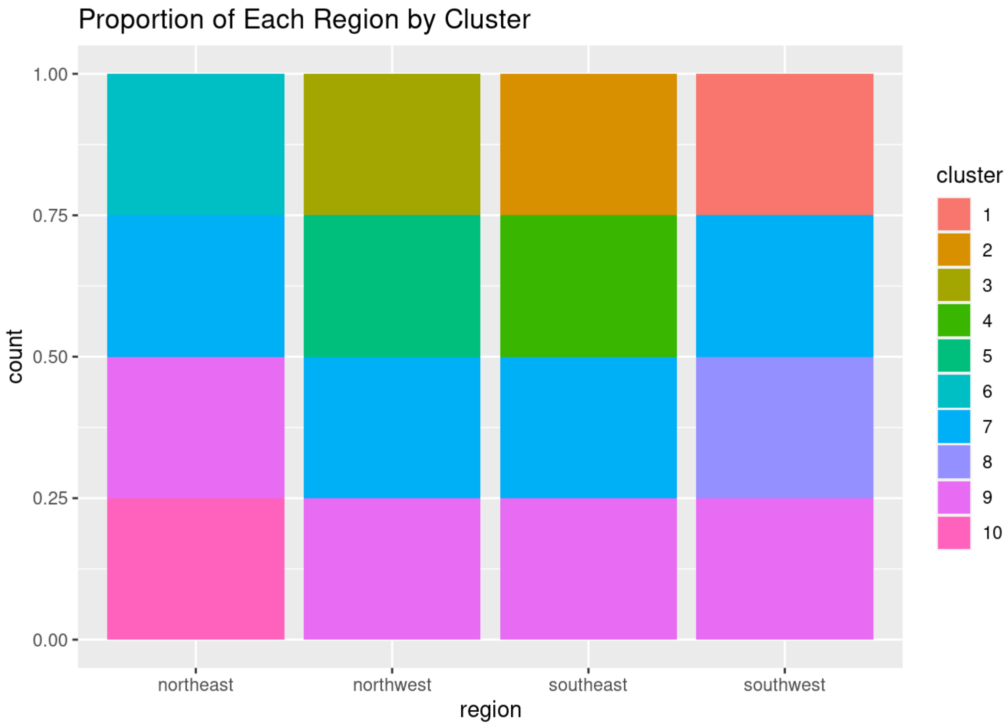
```
#proportion of smoker or non smokers in each cluster
insurance %>%
  mutate(cluster = as.factor(insurance_gpam$clustering)) %>%
  group_by(cluster, smoker) %>%
  summarize(freq = n()) %>%
  ggplot(aes( x = smoker, fill = cluster)) +
  geom_bar(position = 'fill') +
  labs(title = 'Proportion of Smoker or Non Smokers in each Cluster')
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```



```
#proportion of each region by cluster
insurance %>%
  mutate(cluster = as.factor(insurance_gpam$clustering)) %>%
  group_by(cluster, region) %>%
  summarize(freq = n()) %>%
  ggplot(aes( x = region, fill = cluster)) +
  geom_bar(position = 'fill') +
  labs(title = 'Proportion of Each Region by Cluster')
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```



Utilizing the Gower Pam Clustering techniques the medians of the variables by the ten groups clustered by the function are:

```
#cat and numeric summary of each cluster for pam
insurance[insurance_gpam$id.med,]
```

```
## # A tibble: 10 × 7
##      age sex      bmi children smoker region    charges
##    <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
##  1    42 female  29          1 no     southwest   7051.
##  2    41 male    34.2        1 no     southeast   6290.
##  3    40 male    29.4        1 no     northwest   6394.
##  4    38 female  30.7        1 no     southeast   5977.
##  5    41 female  28.3        1 no     northwest   7154.
##  6    38 male    28.0        1 no     northeast   6067.
##  7    48 female  33.1        0 yes    southeast  40974.
##  8    43 male    30.1        1 no     southwest   6849.
##  9    36 male    35.2        1 yes    southeast  38709.
## 10    38 female  27.3        1 no     northeast   6555.
```

From these clusters we can see that the highest insurance charges are in a group characterized as 48 year old females that do smoke in the southeast region. The highest factor seeming to affect boundary between low and high classifications of insurance charges is the smoking status as the only two clusters with yes smoking status are significantly different from the other eight clusters at around $39,000 dollars.

Discussion

Are the age and body mass index of a primary beneficiary member significant predictors of the individual medical costs billed by health insurance companies? By no means was this project easy, however the challenge itself highlighted the importance of good, clean and correlative data. Based upon our research we witness that individuals who are higher in age are at risk of experience higher insurance charges. Our evidence for this can be attributed to Fig 1.1 In which we explore the relationship between Age and Charges and realize that there is an evidence of a relationship. Although it contributes to greatly to our first dimension with a value of 35.1 percent, its evident that is not a good predictor of individual medical costs billed by insurance companies being that there are similar value points between all clusters for Age and BMI exhibits little contribution to the first dimension explaining the variance of our variables with our data.

Is smoking status, regional area, and number of children of the beneficiary member on an insurance plan effective classifiers for the medical costs billed by the health insurance companies? To effectively classify the medical costs utilizing the variables smoker status, region, and number of children on the insurance plan through a logistic regression model has produced less than effective results. Using the root square mean value to discern the error margin a large value of approximately $7,500 produce, identifying this limitation significantly skewed predicted charges that affect accurate classification. Creating PAM clustering with the categorical variables using Gower's dissimilarities hints this idea of smoking status being the prime significant effect in classification by the only two ‘yes’ clusters by having well over average billed health insurance charges. Despite the limitations of the model, some discernment of the effect of the variables can be deciphered knowing that smoking status is significant and number of children having some slight impact while region appears insignificant.

Acknowledgements

With each group member having their respective questions, the work was divided equally among both collaborators. Each research question required independent analysis of data exploration, model creation, clustering, and discussion of findings. Special thanks to Professor Guyot and the TA for providing and being the resources needed to complete the project.