Initial Code and Results

##Load Dataset into R

```
insurance <- read.csv("C:/Users/Ilan/Desktop/insurance.csv")
```

Load the needed libraries

```
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(xgboost)
```

```
##
## Attaching package: 'xgboost'

## The following object is masked from 'package:dplyr':
##
##     slice
```

```
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:psych':
##
##     outlier

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
(tinytex.verbose = TRUE)
```

```
## [1] TRUE
```

##A quick summary of the dataset reveals the different attributes we are dealing with ##is.na reveals if there are any missing variables in the dataset

```
summary(insurance)
```

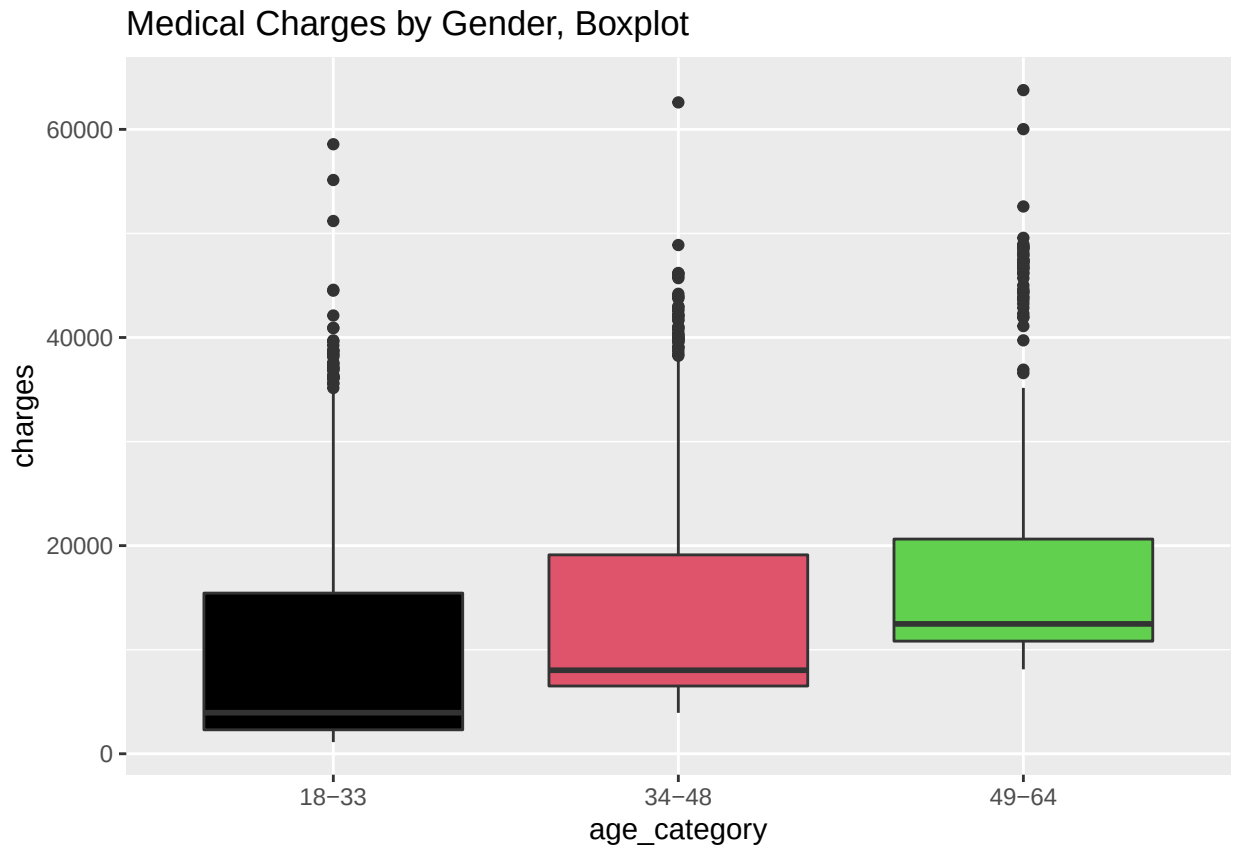
```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region             charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

```
colSums (is.na(insurance))
```

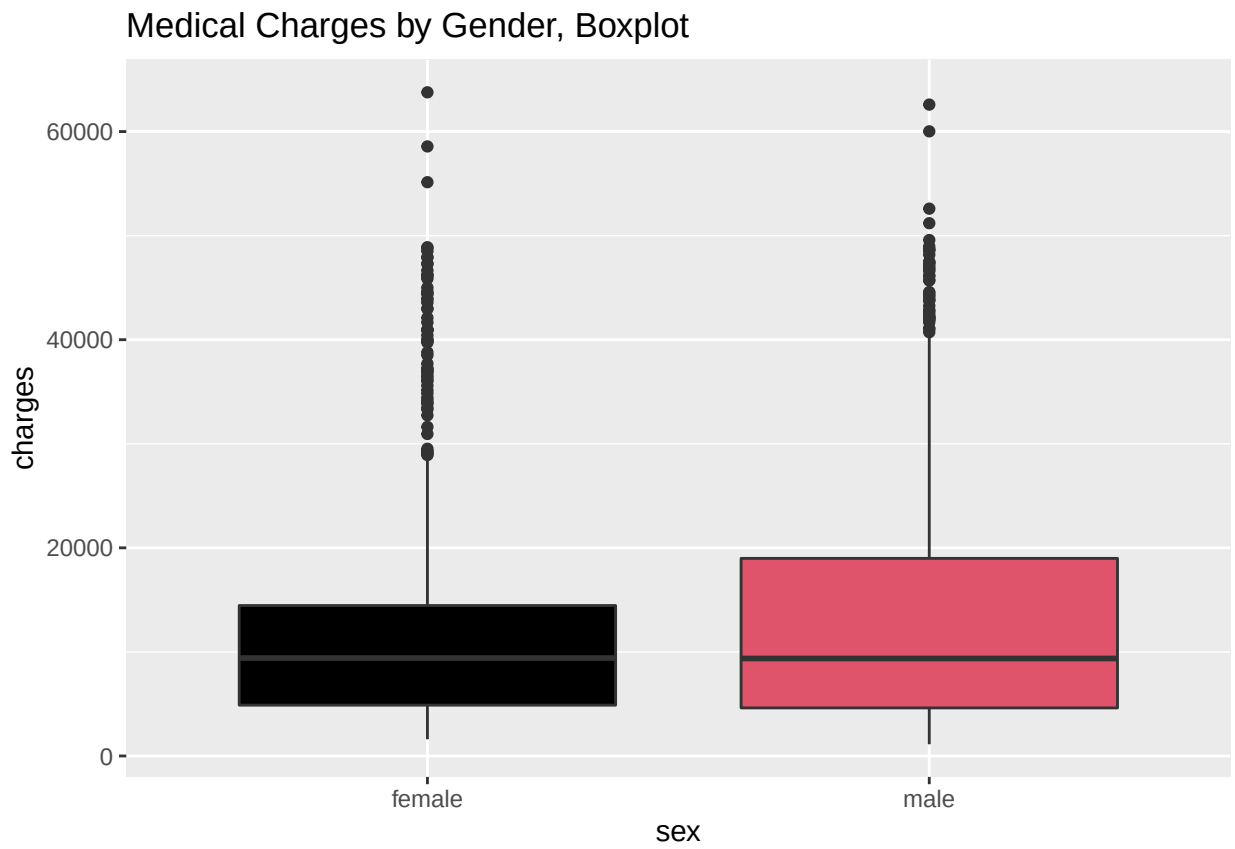
```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        0
```

##BOXPLOTS

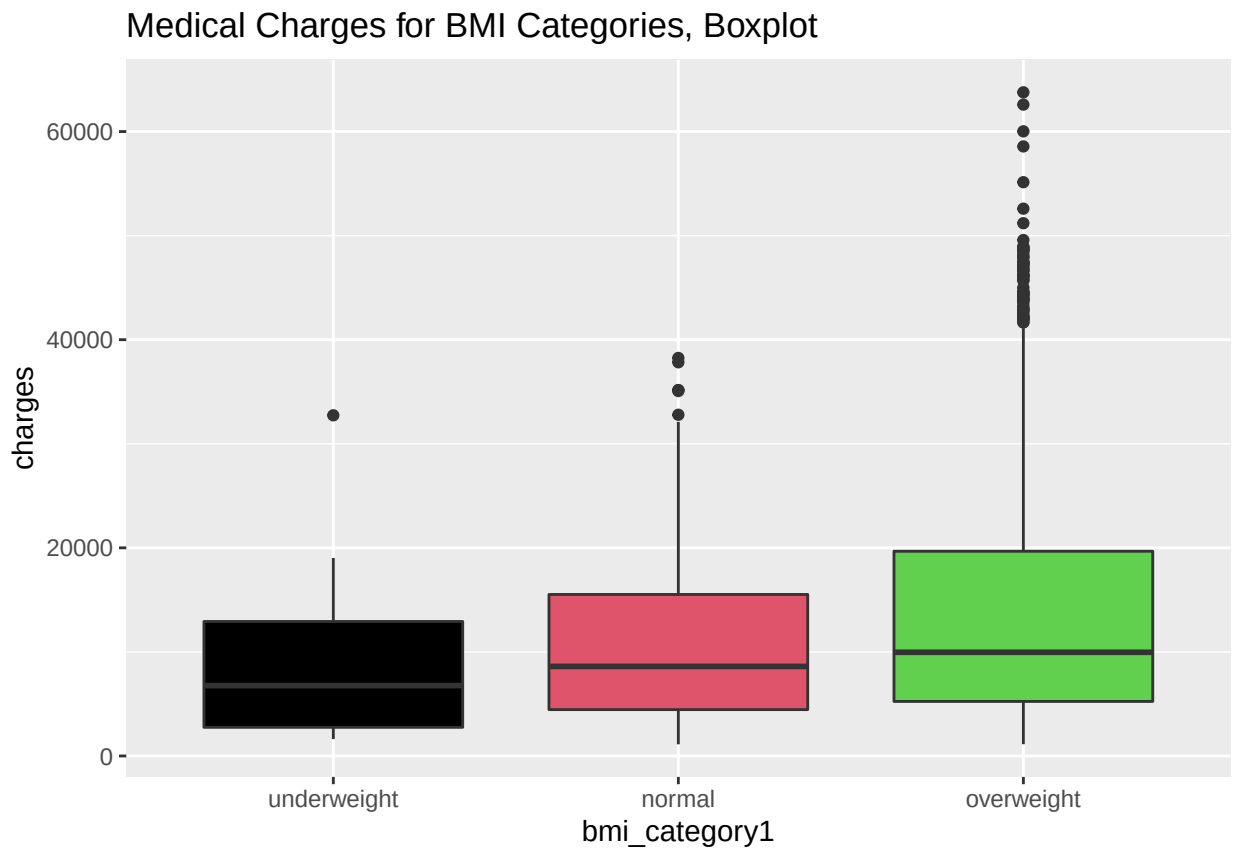
```r
##A study was done about age to disease correlation. K-means was used to cluster the age groups, thus t
for(i in 1:nrow(insurance)){
  if(insurance$age[i] < 34){
    insurance$age_category[i] = "18-33"
  }else if(insurance$age[i] > 33 & insurance$age[i] < 49){
    insurance$age_category[i] = "34-48"
  }else{
    insurance$age_category[i] = "49-64"
  }
}

#In this age boxploto we are able to see that as age groups increase the charges go up as indivduals ar
ggplot(data = insurance, aes(age_category, charges)) + geom_boxplot(fill = c(1:3)) + ggtitle("Medical C
```



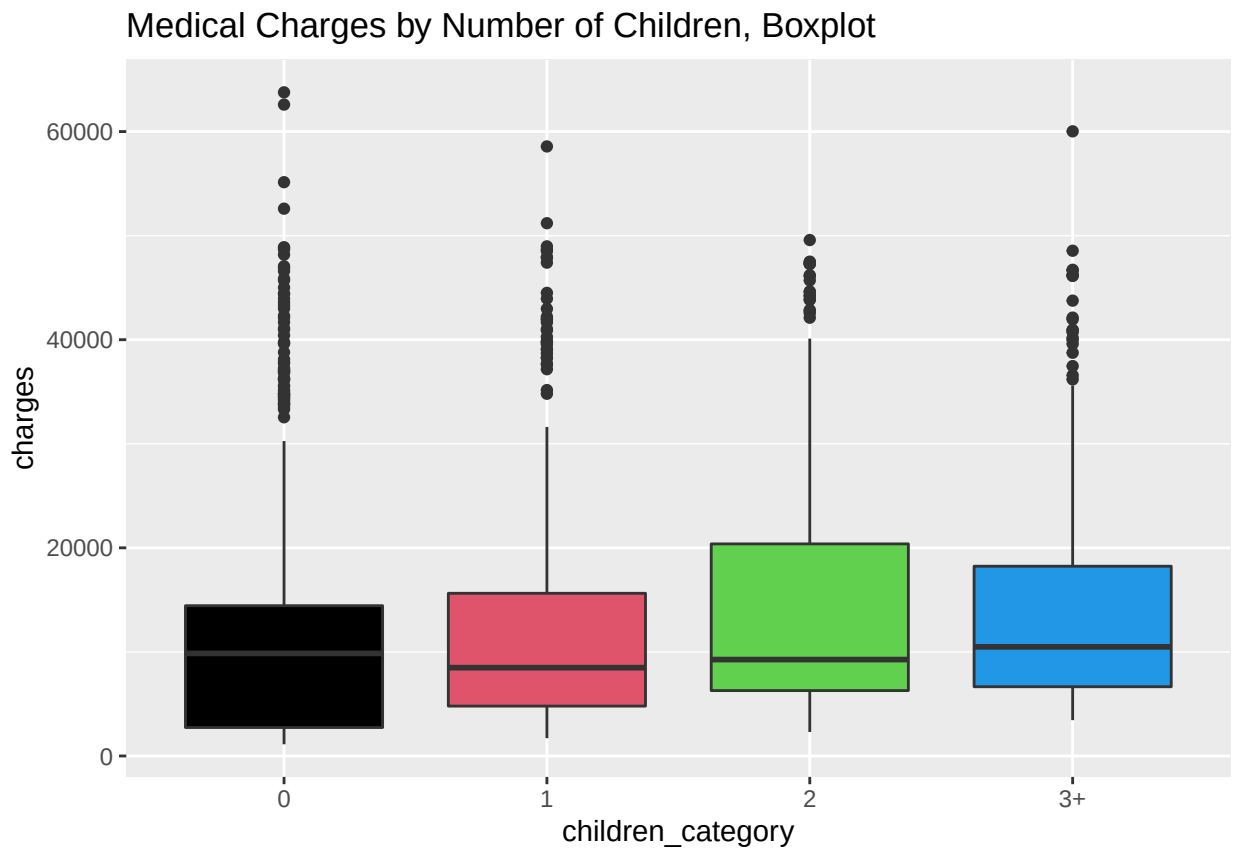

```r
##For gender in relation to charges we are able to see that the data is relatively similar to and doesn
ggplot(data = insurance, aes(sex, charges)) + geom_boxplot(fill = c(1:2)) + ggtitle("Medical Charges by
```



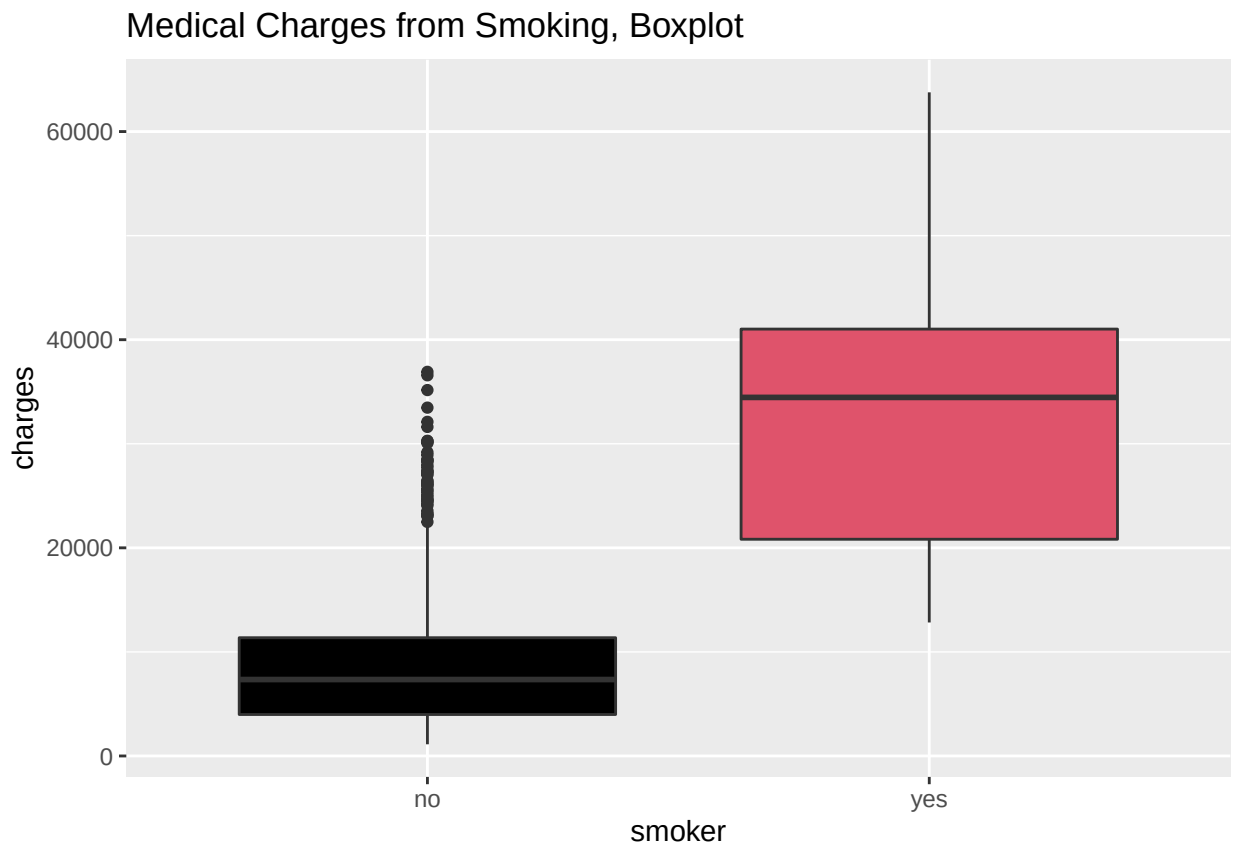

```
##For BMI it was decided to split the data into 3 categories to allow  the data to be more useful and h
for (j in 1:nrow(insurance)) {
  if(insurance$bmi[j] < 18.5){
    insurance$bmi_category[j] = "underweight"
  }else if(insurance$bmi[j] > 30){
    insurance$bmi_category[j] = "overweight"
  }else{
    insurance$bmi_category[j] = "normal"
  }
}

##ggplot2 organizes the boxplot based on alphabetical order, here we are creating a new variable that o
insurance$bmi_category1 <- factor(insurance$bmi_category, levels=unique(as.character(insurance$bmi_categ

##In this boxplot of BMI in relation to charges we are able to see that being overweight has a major im
ggplot(data = insurance, aes(bmi_category1, charges)) + geom_boxplot(fill = c(1:3)) + ggtitle("Medical C
```



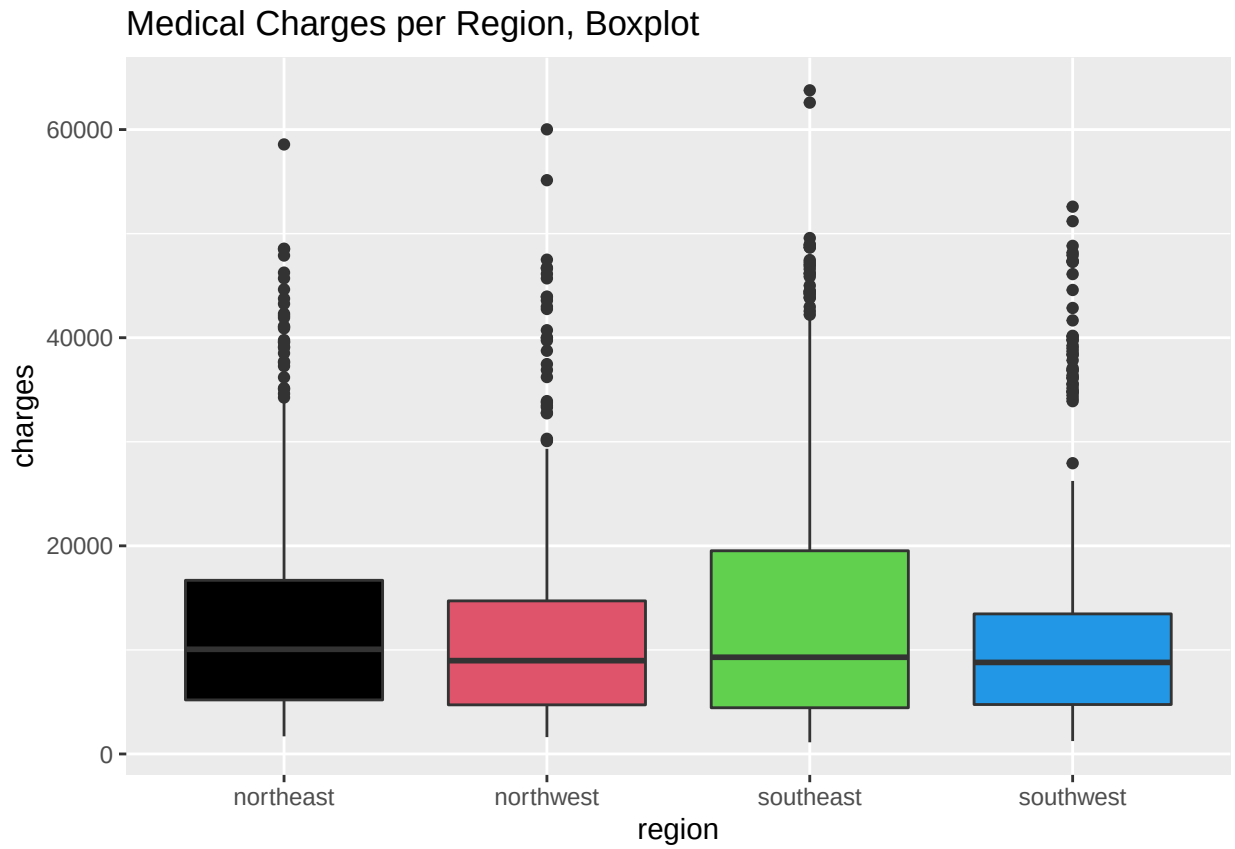

```
##Here the children attribute was categorized into 3 different parts in order to create a balanced dist
for (k in 1:nrow(insurance)){
  if(insurance$children[k] == 0){
    insurance$children_category[k] = "0"
  }else if(insurance$children[k] == 1){
    insurance$children_category[k] = "1"
  }else if(insurance$children[k] == 2){
    insurance$children_category[k] = "2"
  }else{
    insurance$children_category[k] = "3+"
  }
}

##In this boxplot we are able to see that the number of children doesnt have a drastic affect on charge
ggplot(data = insurance, aes(children_category, charges)) + geom_boxplot(fill = c(1:4)) + ggtitle("Medi
```




```
##In this boxplot for smoking we are able to see that individuals who smoke have higher charges than non
ggplot(data = insurance, aes(smoker, charges)) + geom_boxplot(fill = c(1:2)) + ggtitle("Medical Charges
```

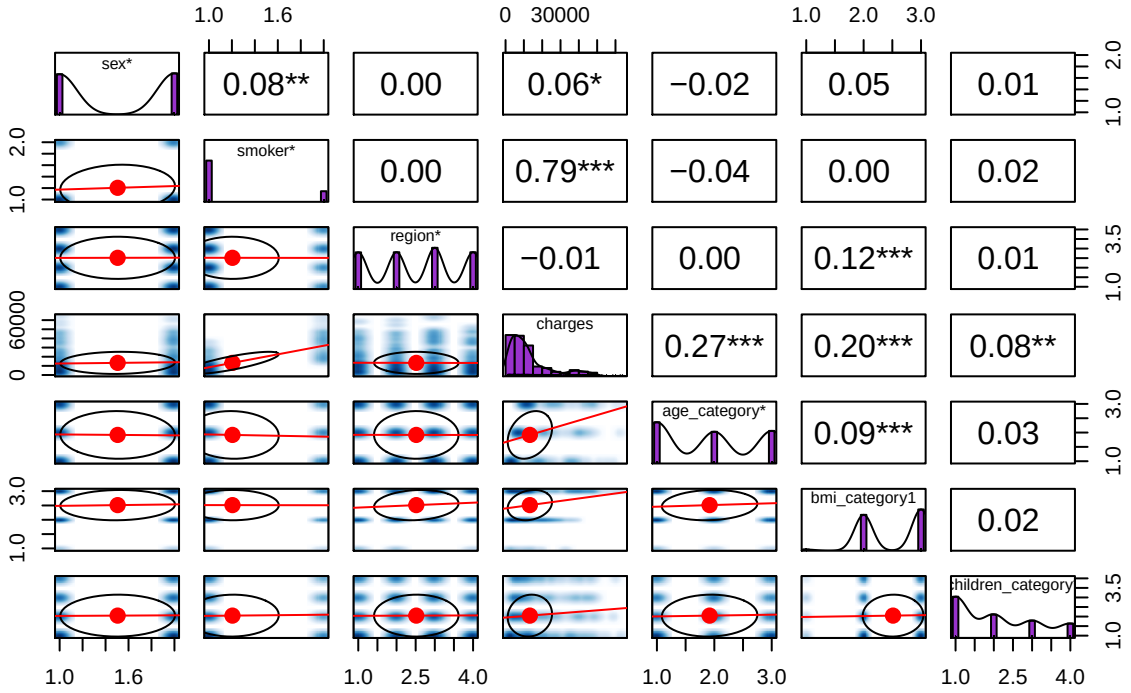



```
##For the region boxplot it is evident that there is much affect on medical charges and will most likel
ggplot(data = insurance, aes(region, charges)) + geom_boxplot(fill = c(1:4)) + ggtitle("Medical Charges
```



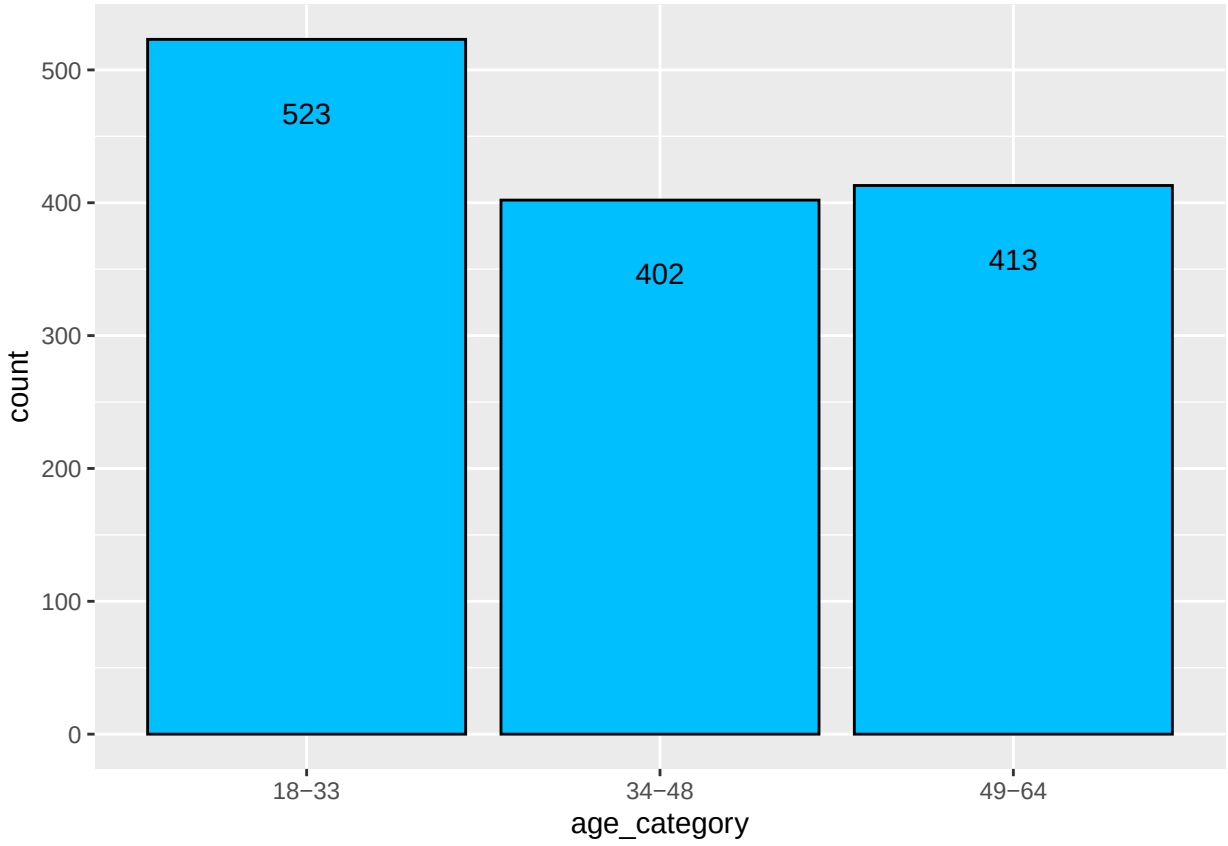
##CORRELATION MATRIX

```
##Created a new data frame in order to drop the original attributes that have been altered into new one
insurance_new <- select(insurance, -c(age, bmi, bmi_category, children))

##Within the correlation matrix it is evident that there are 3 main attributes that are correlated with
pairs.panels(insurance_new, pch = 1, lm = TRUE, cex.cor = 1, hist.col="darkorchid", smoother = T, show.
             density = TRUE, stars = TRUE, main="Correlation Matrix")
```

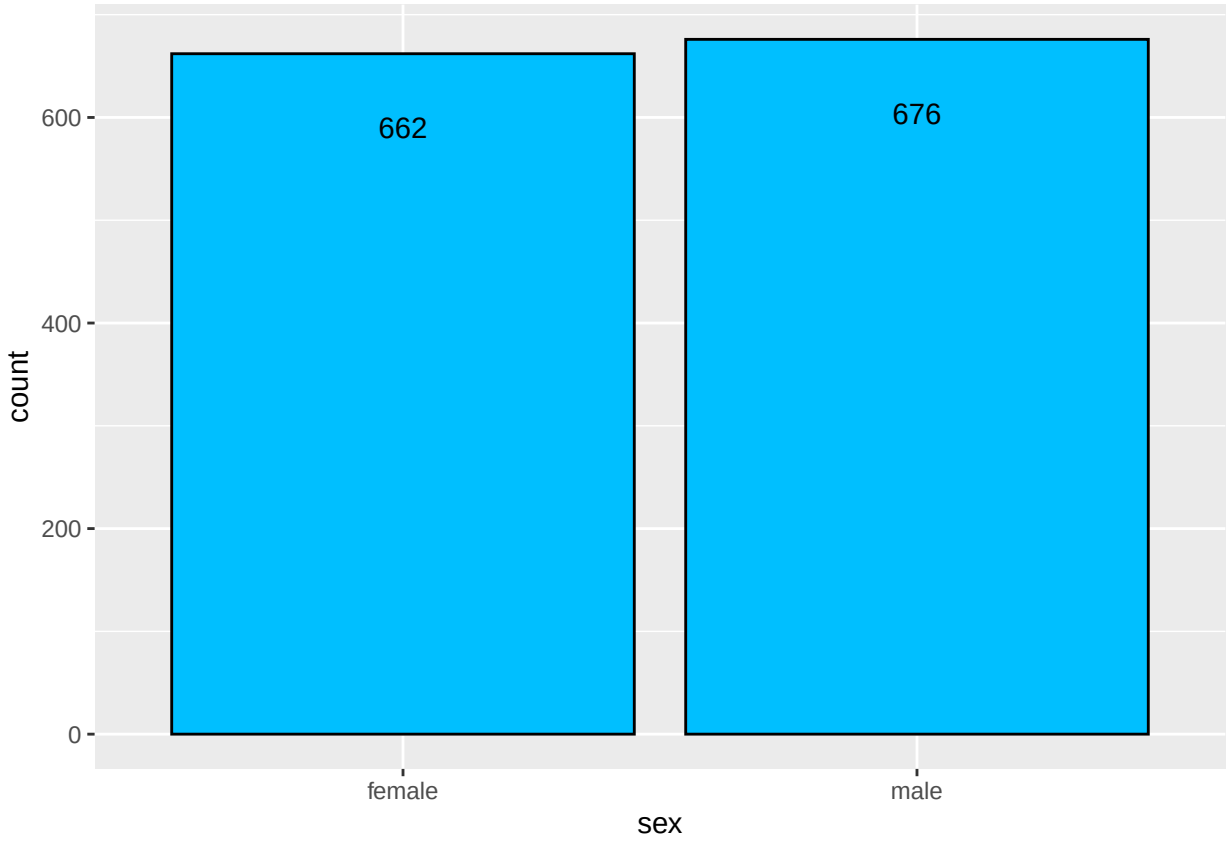


##HISTOGRAMS

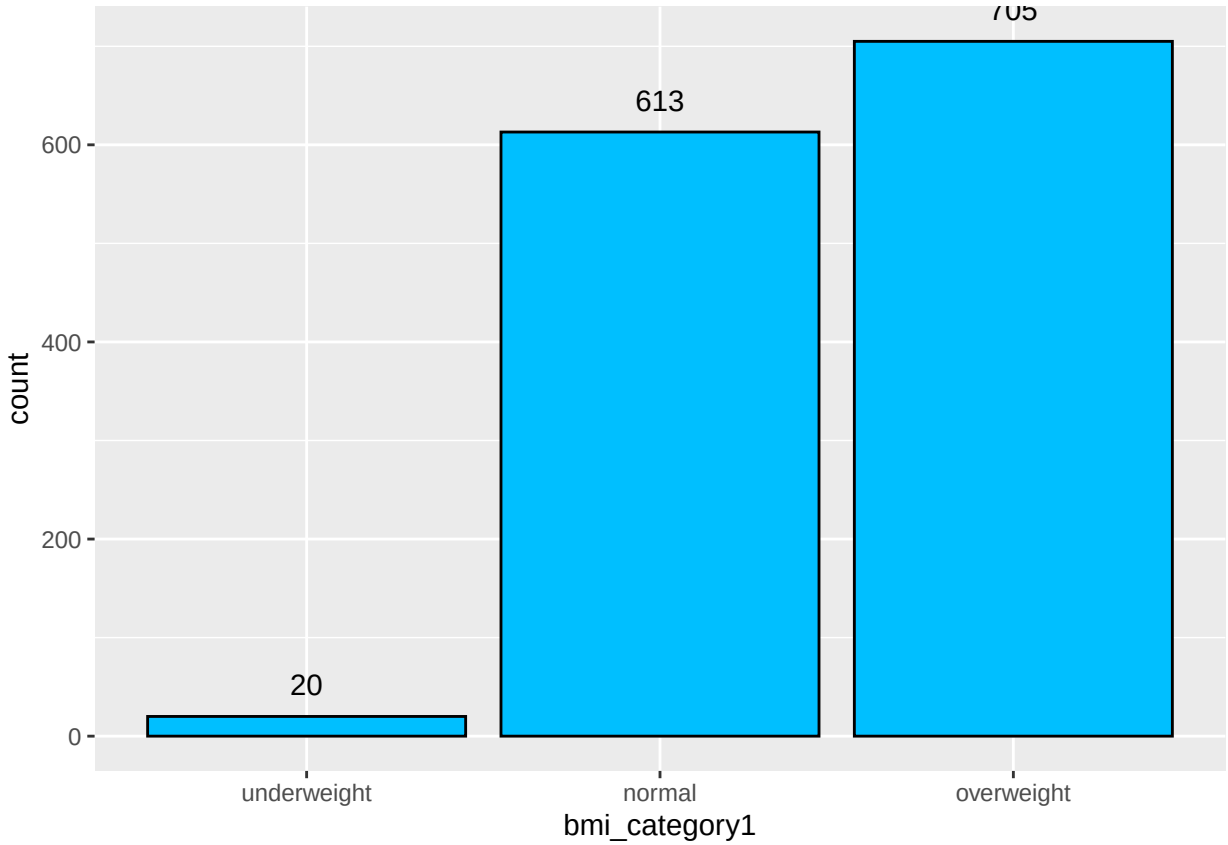
```
##This histogram is used to see the distribution of age to make sure it is balanced
ggplot(insurance, aes(x = age_category)) + geom_bar(color = "black", fill = "deepskyblue") + geom_text(s
```



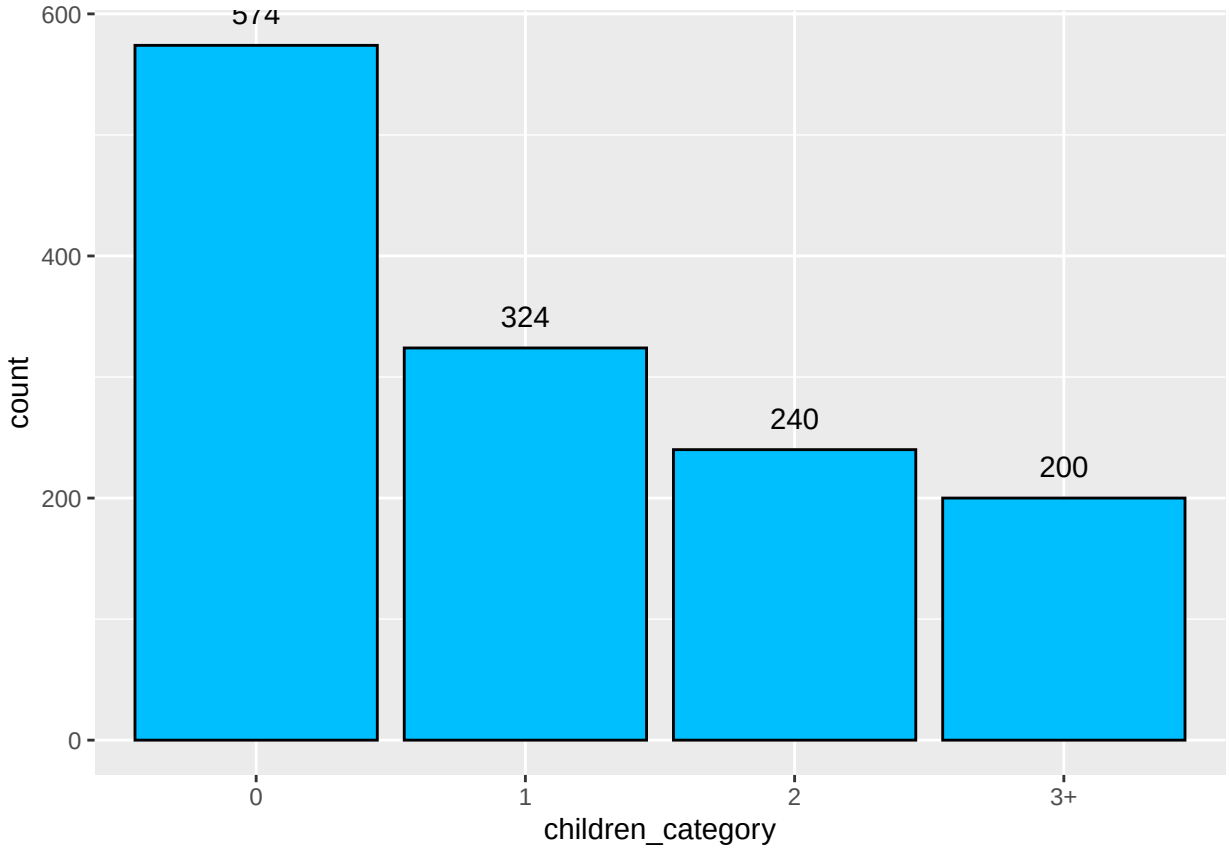

```
##This histogram is used to see the distribution of sex to make sure it is balanced
ggplot(insurance, aes(x = sex)) + geom_bar(color = "black", fill = "deepskyblue") + geom_text(stat="cou
```



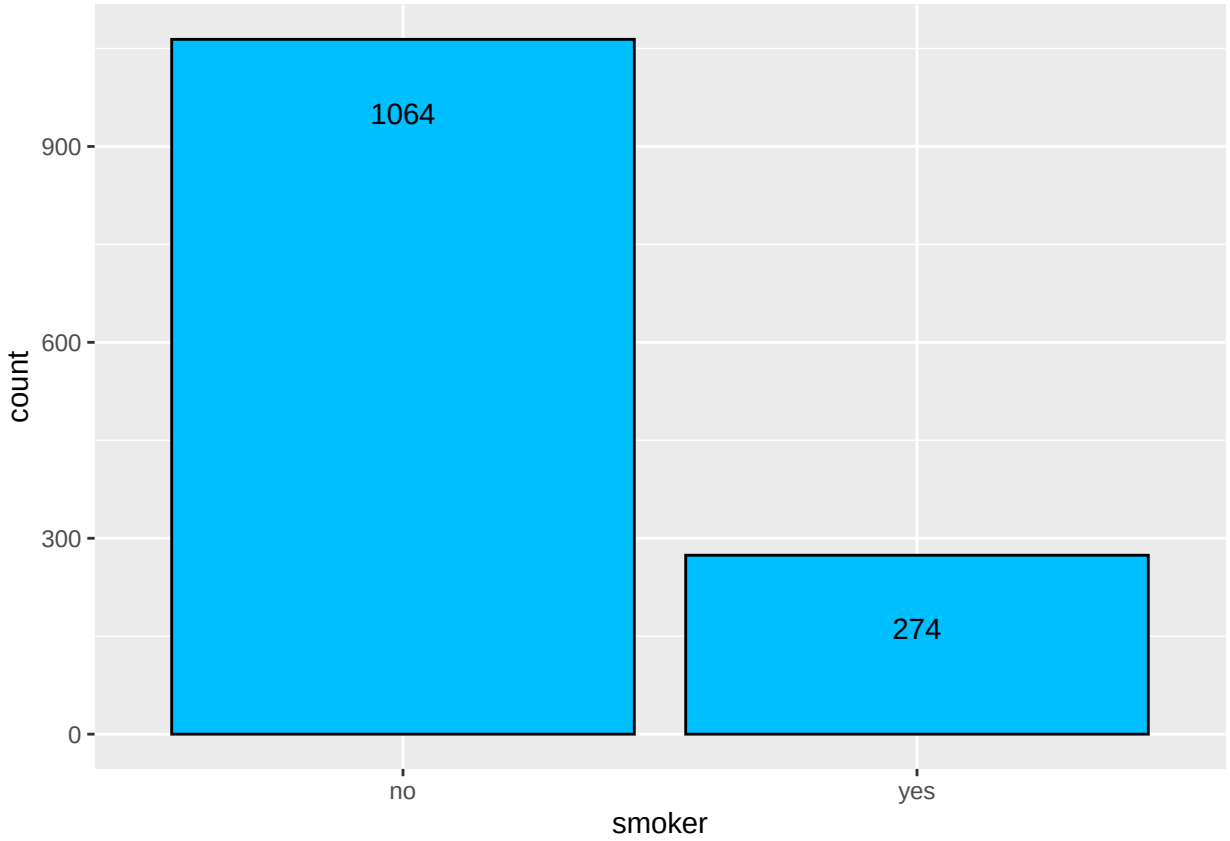

```
##Here one is able to see that the underweight category has a small amount of data assigned  to it, dur
ggplot(insurance, aes(x = bmi_category1)) + geom_bar(color = "black", fill = "deepskyblue")+ geom_text(s
```



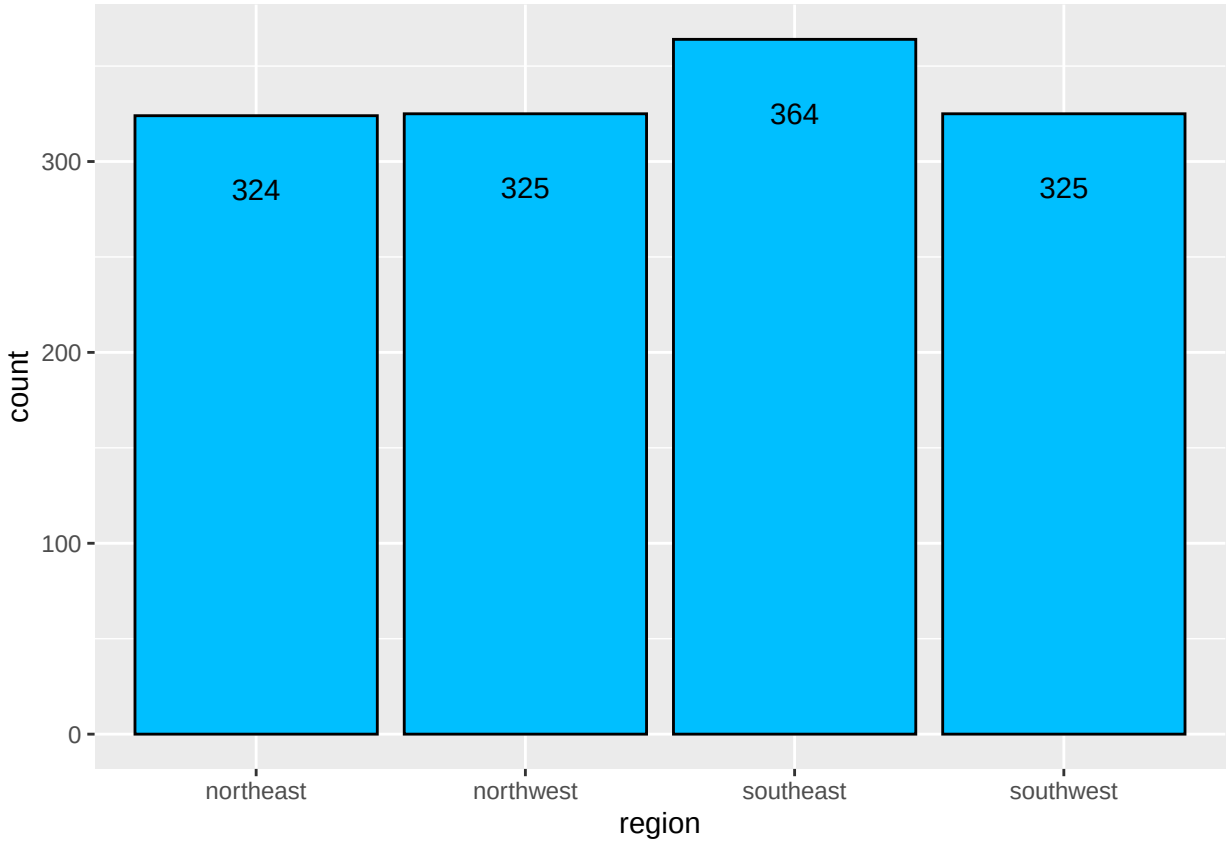

```
##In this histogram we are checking the distribution of the amount of children for clients
ggplot(insurance, aes(children_category)) + geom_bar(color = "black", fill = "deepskyblue") +
  geom_text(stat="count", aes(label=..count..), vjust=-1)
```




```
##In this histogram we can see the distribution of smokers is not similar but since the correlation mat
ggplot(insurance, aes(x = smoker)) + geom_bar(color = "black", fill = "deepskyblue") + geom_text(stat="
```

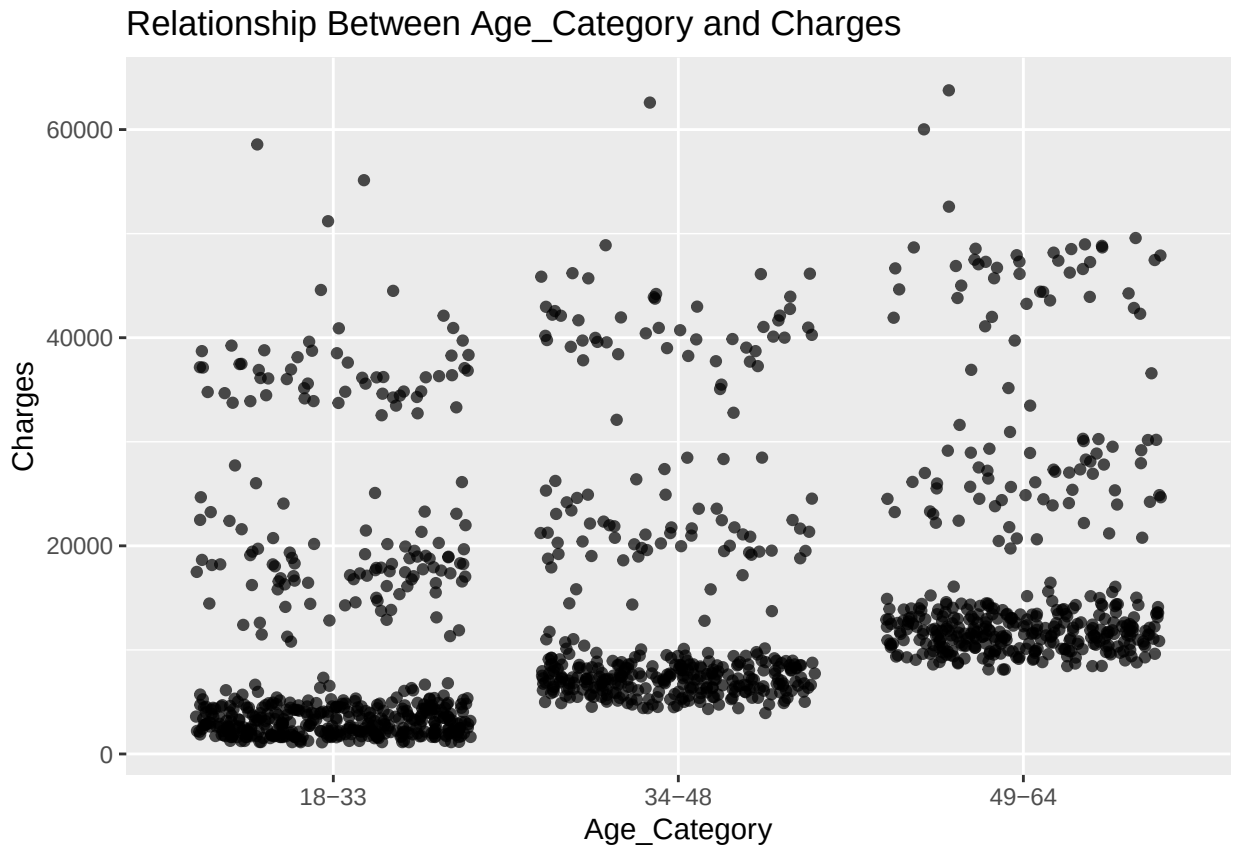



```
##The distribution for region is very close and will not be altered
ggplot(insurance, aes(x = region)) + geom_bar(color = "black", fill = "deepskyblue") + geom_text(stat="
```

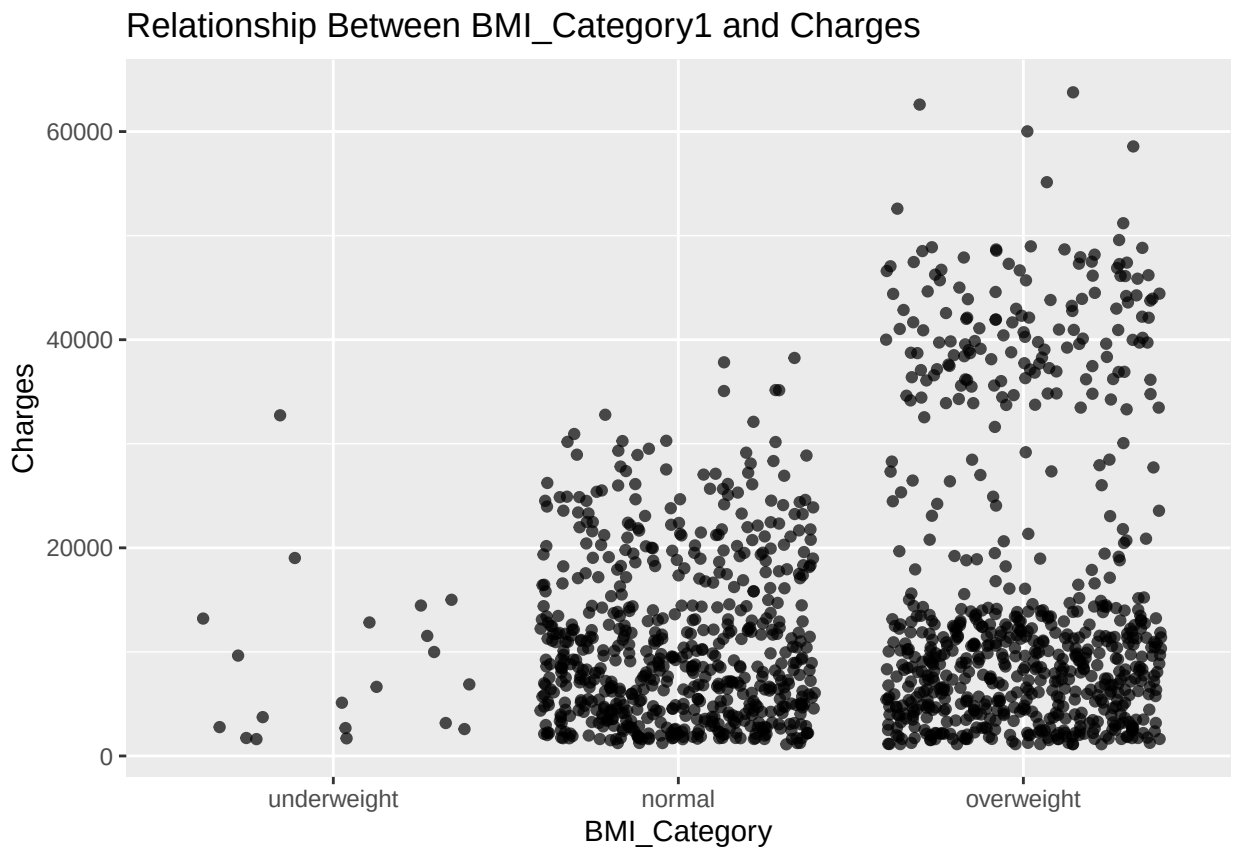


##SCATTER PLOTS

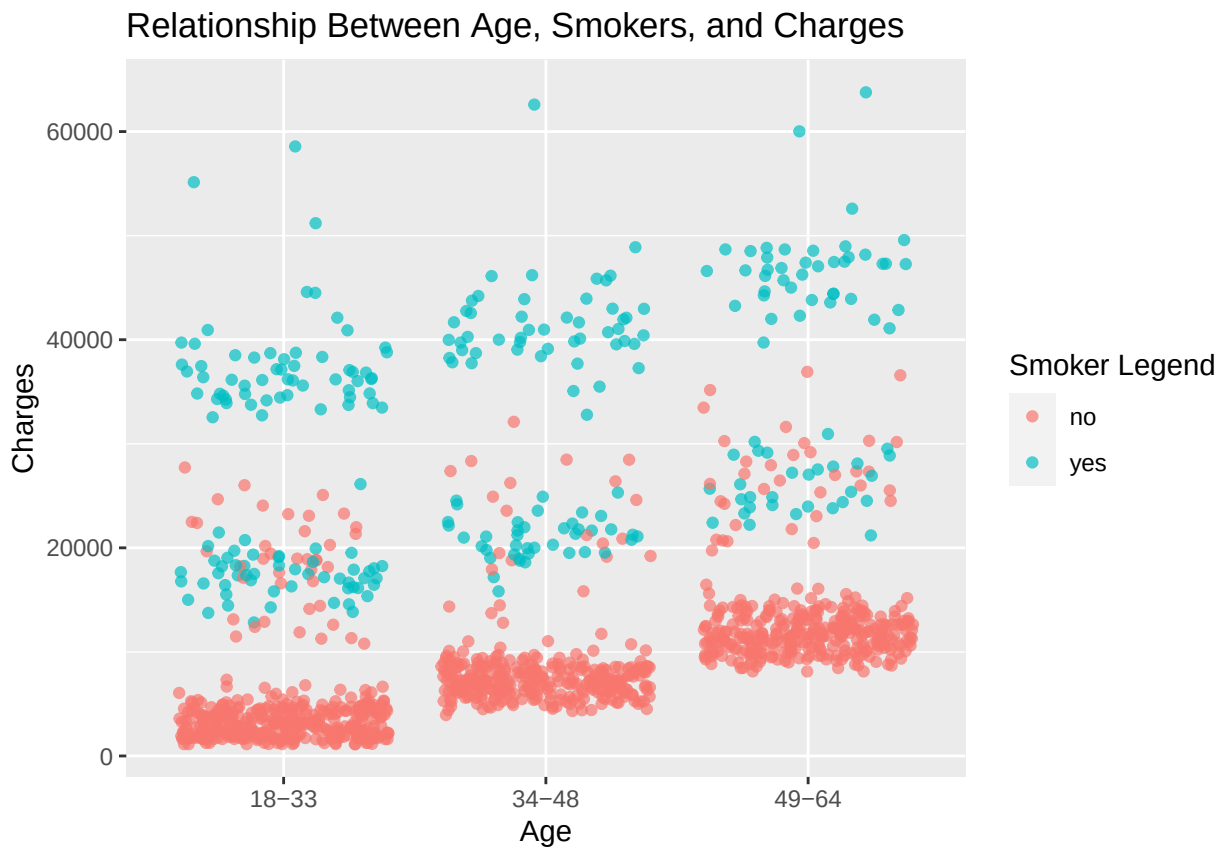
```
##In this scatter plot of age_category we can see that each age group is split up into 3 clusters, this
ggplot(insurance, aes(x = age_category, y = charges)) + geom_jitter(aes(age_category), alpha = 0.7) + la
```



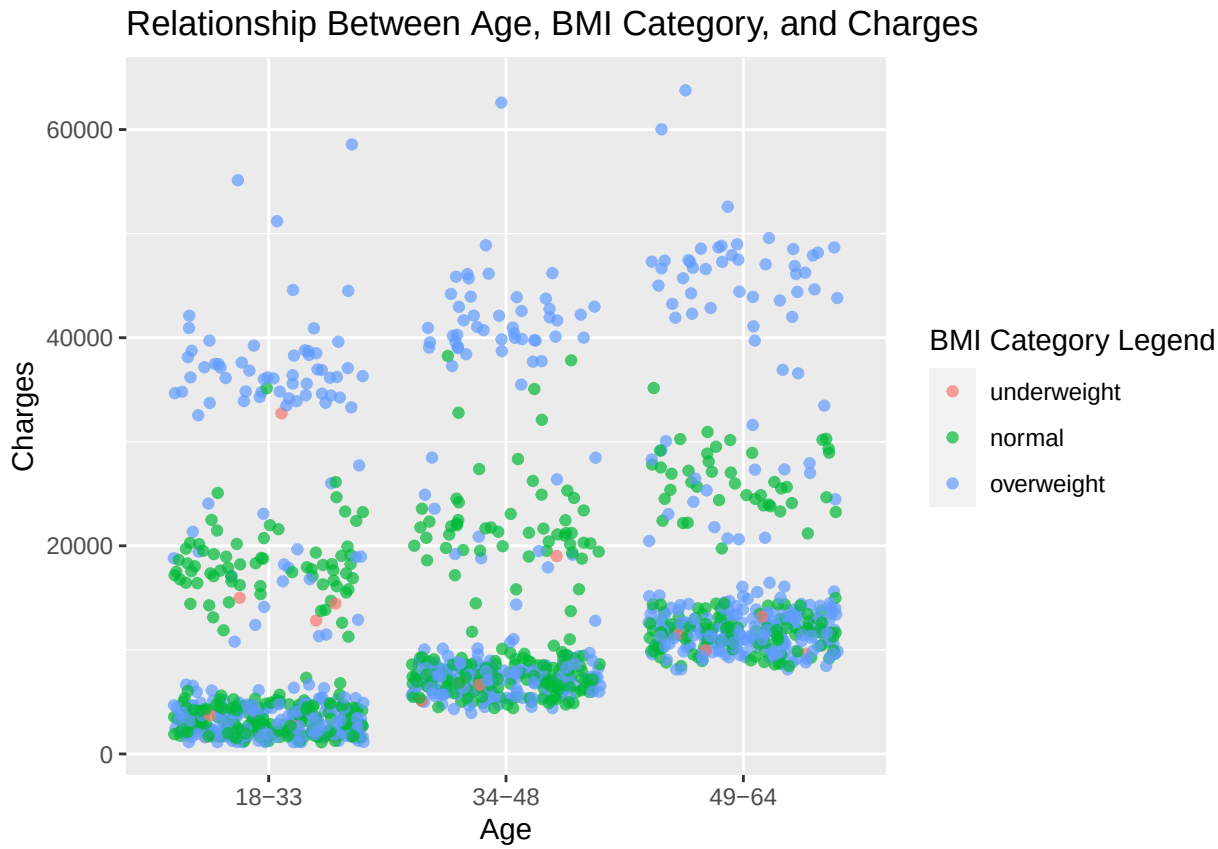

```
##In this scatter plot of BMI category we can see that overweight category has higher charges than norm
ggplot(insurance, aes(x = bmi_category1, y = charges)) + geom_jitter(aes(bmi_category1), alpha = 0.7) +
  labs(x = "BMI_Category", y = "Charges") +  ggtitle("Relationship Between BMI_Category1 and Charges")
```



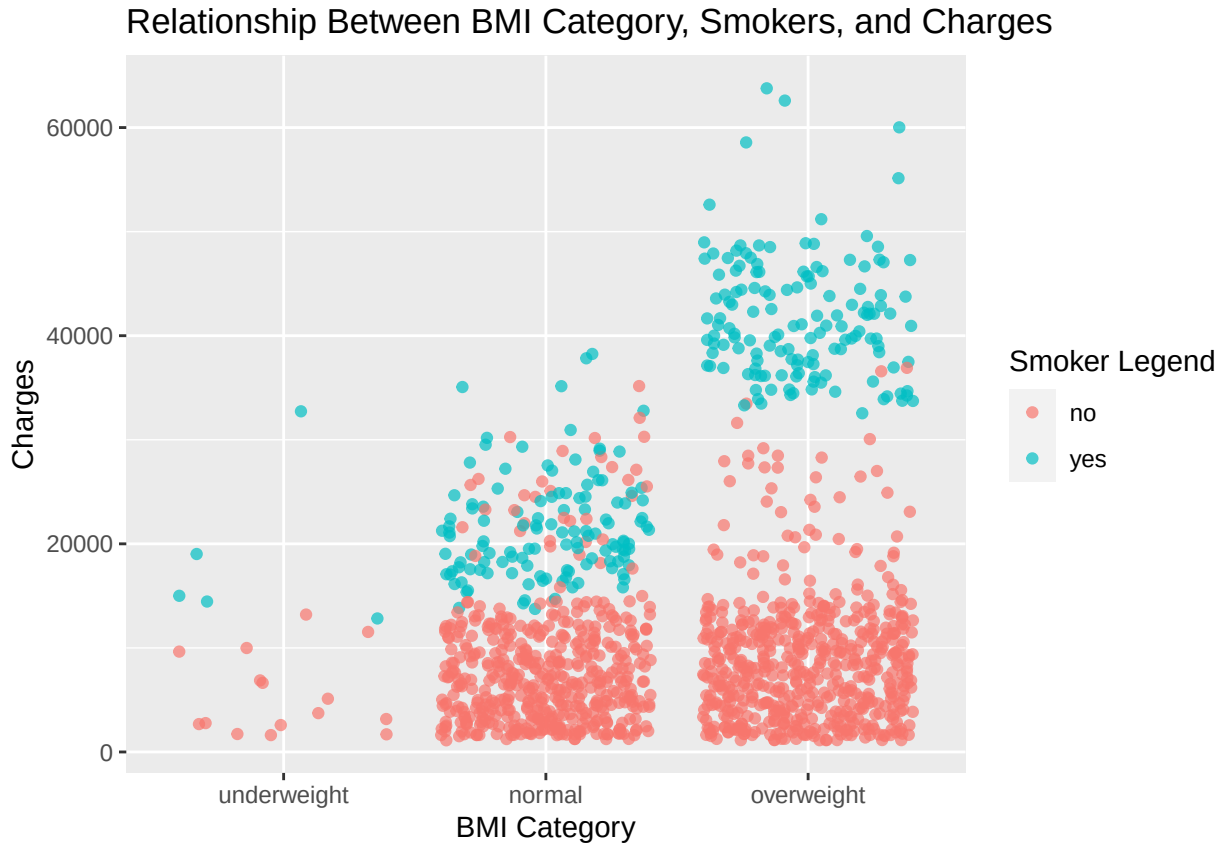

```
##In this scatter plot we can see the age distribution and how smoking affects the charges. As you can
ggplot(insurance, aes(x = age_category, y = charges)) + geom_jitter(aes(color = smoker), alpha = 0.7) +
  labs(x = "Age", y = "Charges", col = "Smoker Legend") +  ggtitle("Relationship Between Age, Smokers, a
```




```
##Here we can see that BMI has a a big affect on the charges as a majority of the overweight individuals
ggplot(insurance, aes(x = age_category, y = charges)) + geom_jitter(aes(color = bmi_category1), alpha =
  labs(x = "Age", y = "Charges", col = "BMI Category Legend") +  ggtitle("Relationship Between Age, BMI
```




```
##Here we can see that if you are obese and a smoker your charges skyrocket and you can see the major g
ggplot(insurance, aes(x = bmi_category1, y = charges)) + geom_jitter(aes(color = smoker), alpha = 0.7) +
  labs(x = "BMI Category", y = "Charges", col = "Smoker Legend") +  ggtitle("Relationship Between BMI Ca
```



##ALGORITHMS

```
##MULTI-LINEAR REGRESSION

##Setting the seed allows me to reproduce the output of the algorithm
set.seed(1)

##Here I am splitting the data into a training set and test set
selection <- sample(1:nrow(insurance_new), 0.8 * nrow(insurance_new))
train_LM = insurance_new[selection, ]
test_LM = insurance_new[-selection, ]

##This is the first model with all the attributes
LM <- lm(charges ~ sex + smoker + region + age_category + bmi_category1 + children_category, data = tra
##First off we see that the Multiple R-squared values is 0.7729 which shows that the model explains the
##Furthermore we can analyze the p-values to get a better understanding of which variable is helping th
summary(LM)
```

```
##
## Call:
## lm(formula = charges ~ sex + smoker + region + age_category +
##     bmi_category1 + children_category, data = train_LM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12826.9  -3648.2   -631.3   1762.9  30181.4
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -257.7     1474.9  -0.175 0.861323
## sexmale                    -97.1      371.6  -0.261 0.793924
## smokeryes                23395.6      464.9  50.324  < 2e-16 ***
## regionnorthwest           -127.9      529.6  -0.241 0.809221
## regionsoutheast           -482.0      520.6  -0.926 0.354699
## regionsouthwest           -403.7      535.2  -0.754 0.450890
## age_category34-48         3411.2      460.8   7.403 2.71e-13 ***
## age_category49-64         8556.2      449.6  19.030  < 2e-16 ***
## bmi_category1normal       2599.2     1456.7   1.784 0.074662 .
## bmi_category1overweight   6441.5     1463.7   4.401 1.19e-05 ***
## children_category1         700.0      478.9   1.462 0.144130
## children_category2        1784.0      534.6   3.337 0.000876 ***
## children_category3+       1596.2      558.7   2.857 0.004362 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6028 on 1057 degrees of freedom
## Multiple R-squared:  0.7444, Adjusted R-squared:  0.7415
## F-statistic: 256.6 on 12 and 1057 DF,  p-value: < 2.2e-16
```

```
##XGBOOST
##NOT READY, still testing the paramaters and making sure everything works together
# set.seed(2)
#
# selection1 <- sample(1:nrow(insurance_new), 0.8 * nrow(insurance_new))
# train_XGB = insurance_new[selection1, ]
# test_XGB = insurance_new[-selection1, ]
#
# XGB <- xgboost(data = train_XGB(),
#                label = insurance$charges,
#                eta = 0.1,
#                max_depth = 15,
#                nround = 25,
#                subsample = 0.5,
#                colsample_bytree = 0.5,
#                seed = 2,
#  eval_metric = "merror",
#  objective = "multi:softprob",
#  num_class = 12,
#  nthread = 3
#               )
```

```
##RANDOMFOREST

##Split the data into training, validation, and test
selection2 <- sample(1:nrow(insurance_new), 0.8 * nrow(insurance_new))

training_RF <- insurance_new[selection2, ]
validation_RF <- insurance_new[selection2, ]

##Here we can see that the variance explained is at 83.45% which is very good
rf = randomForest(charges ~ sex + smoker + region + age_category + bmi_category1 + children_category, da
```