Final Project Team 3

Importing and Cleaning Data Set

```
##libraries to run code below
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Library
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 0x0006): Library not loaded: /
##   Referenced from: <BBB44505-4BB3-30FA-9ED6-ABC69D534041> /Library/Frameworks/R.framework/Versions/4
##   Reason: tried: '/opt/X11/lib/libSM.6.dylib' (no such file), '/System/Volumes/Preboot/Cryptexes/OS/
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)),
## stdout = TRUE): running command ''/usr/bin/otool' -L
## '/Library/Frameworks/R.framework/Resources/library/tcltk/libs//tcltk.so'' had
## status 1
```

```
## Could not load tcltk.  Will use slower R code instead.
## Loading required package: RSQLite
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(knitr)
library(shinythemes)
library(shiny)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##orig. data set un-modified
insurance <- read_csv("insurance.csv")
```

```
## Rows: 1338 Columns: 7
## -- Column specification ---------------------------------------------------------
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
##creating modified data frame for use
ins_df <- insurance

##dummy variables and adding to data model
female_dt <- ifelse(ins_df$sex == 'female',1,0)
smoker_dt <- ifelse(ins_df$smoker == 'yes',1,0)
reg_southwest_dt <- ifelse(ins_df$region == 'southwest',1,0)
reg_southeast_dt <- ifelse(ins_df$region == 'southeast',1,0)
reg_northwest_dt <- ifelse(ins_df$region == 'northwest',1,0)
ins_df <- cbind(ins_df,female_dt,smoker_dt,reg_southwest_dt,
                reg_southeast_dt,reg_northwest_dt)
```

Descriptive Statistics for Demographics of the Data Set

```
##averages, ranges, and percentages to understand data
##avg age is 39.21
avg_age <- mean(ins_df$age)
avg_age
```

```
## [1] 39.20703
```

```
##age range is 18 to 64
age_range <- range(ins_df$age)
age_range
```

```
## [1] 18 64
```

```
##avg bmi is 30.66
avg_bmi <- mean(ins_df$bmi)
avg_bmi
```

```
## [1] 30.6634
```

```
##bmi range is 15.96-53.13
bmi_range <- range(ins_df$bmi)
bmi_range
```

```
## [1] 15.96 53.13
```

```
##avg children 1.09
avg_children <- mean(ins_df$children)
avg_children
```

```
## [1] 1.094918
```

```
##children range is 0-5
children_range <- range(ins_df$children)
children_range
```

```
## [1] 0 5
```

```
##avg charges 13270.42
avg_charges <- mean(ins_df$charges)
```

```r
avg_charges
```

```
## [1] 13270.42
```

```r
##charges range is $1,121.87-$63,770.43
charges_range <- range(ins_df$charges)
charges_range
```

```
## [1]  1121.874 63770.428
```

```r
##percent of population female sex .4948
perc_female_sex <- sum(ins_df$female_dt)/nrow(ins_df)
perc_female_sex
```

```
## [1] 0.4947683
```

```r
##percent of population male sex .5052
perc_male_sex <- 1-perc_female_sex
perc_male_sex
```

```
## [1] 0.5052317
```

```r
##percent of population smoker .2048
perc_smoker <- sum(ins_df$smoker_dt)/nrow(ins_df)
perc_smoker
```

```
## [1] 0.2047833
```

```r
##percent of non-smokers .7952
perc_nonsmoker <- 1-sum(ins_df$smoker_dt)/nrow(ins_df)
perc_nonsmoker
```

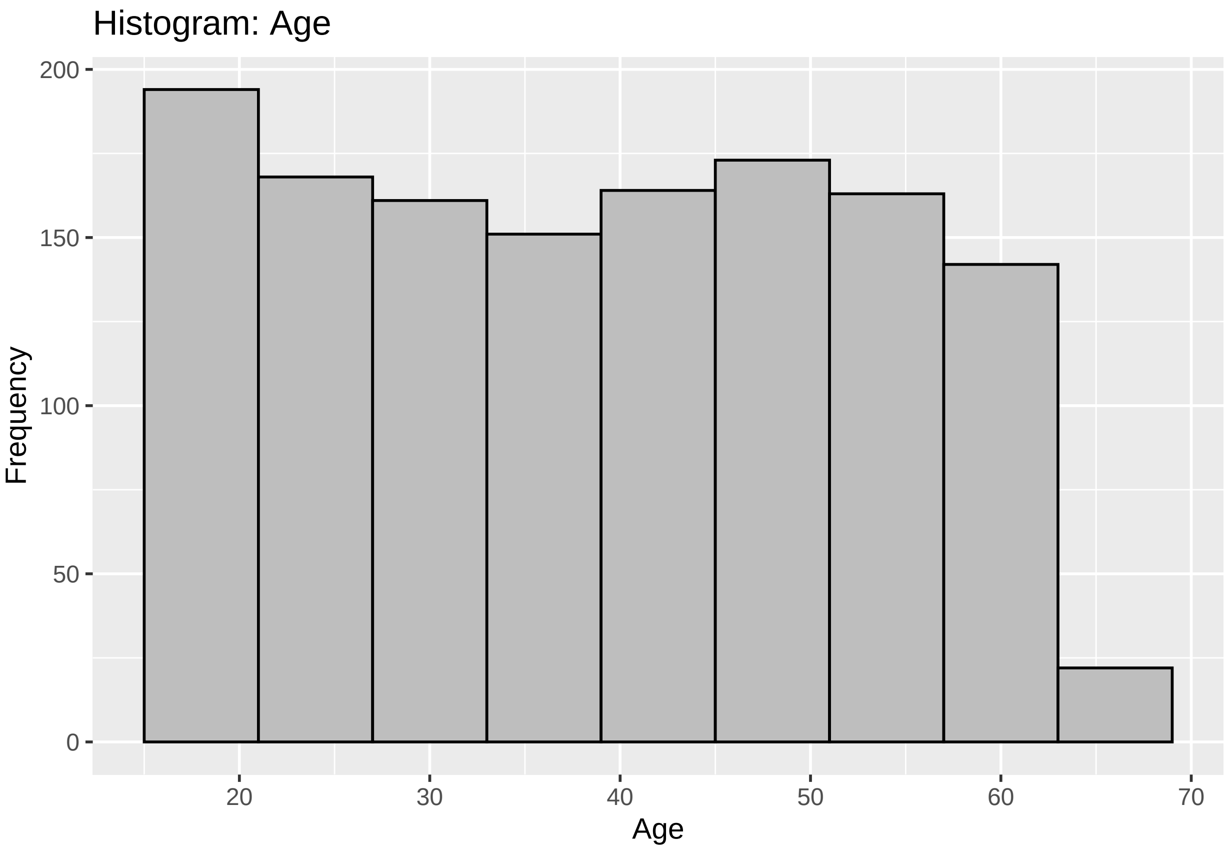
```
## [1] 0.7952167
```

```r
##population by region
regions_descrip <- sqldf("select region, count(*) as region_count from
                         ins_df group by region order by region desc")
regions_descrip
```

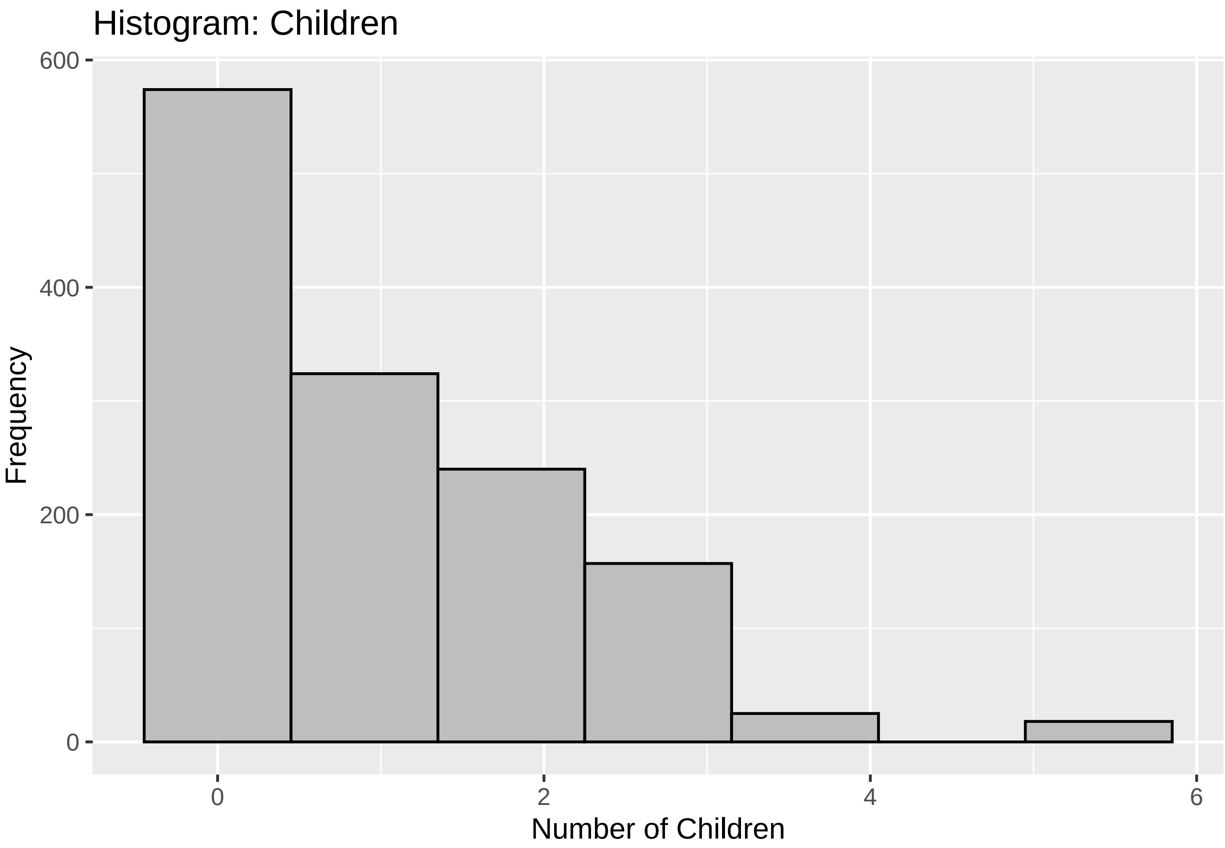
```
##      region region_count
## 1 southwest          325
## 2 southeast          364
## 3 northwest          325
## 4 northeast          324
```

Histograms of Key Demographic Data

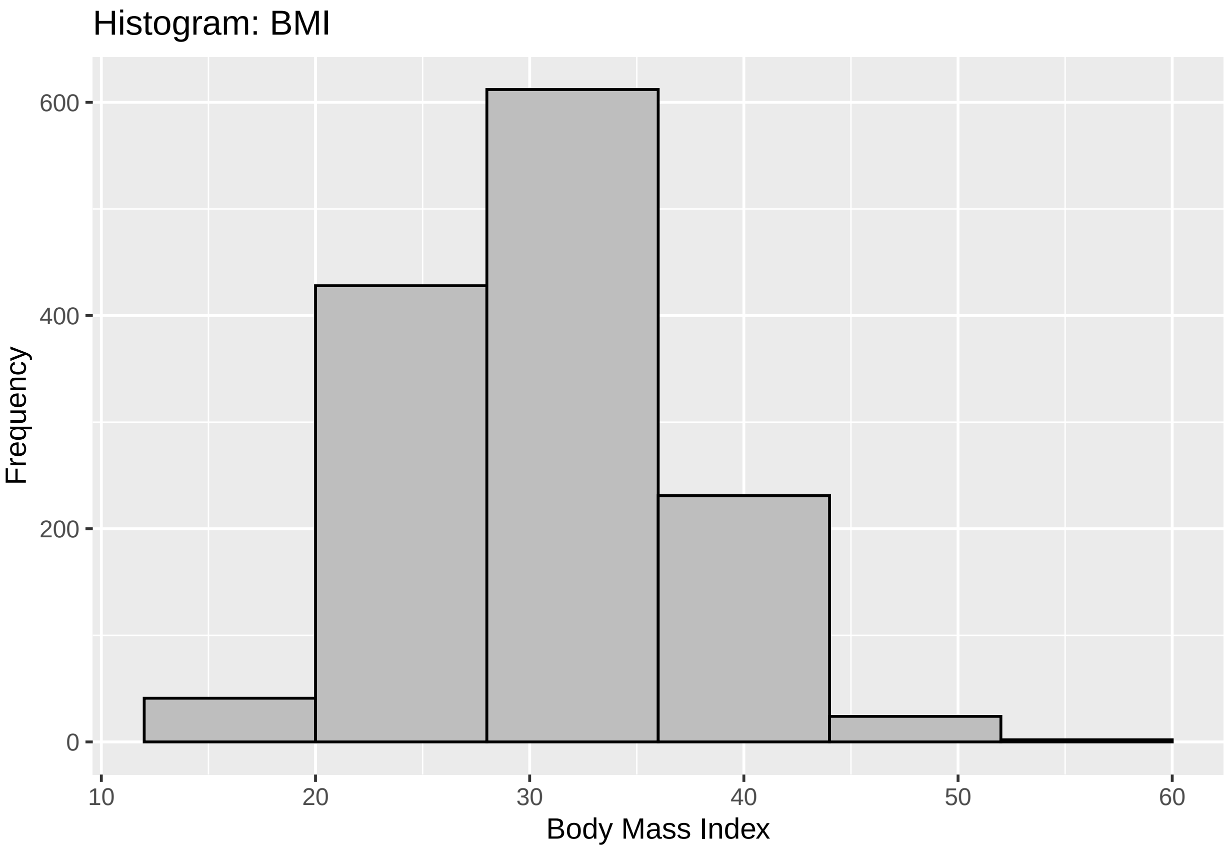
```r
age_hist <- ggplot(data = ins_df, aes(x=age)) +
  geom_histogram(color ="black", fill="grey", binwidth = 6) +
  labs(title = "Histogram: Age", x = "Age", y = "Frequency")
age_hist
```



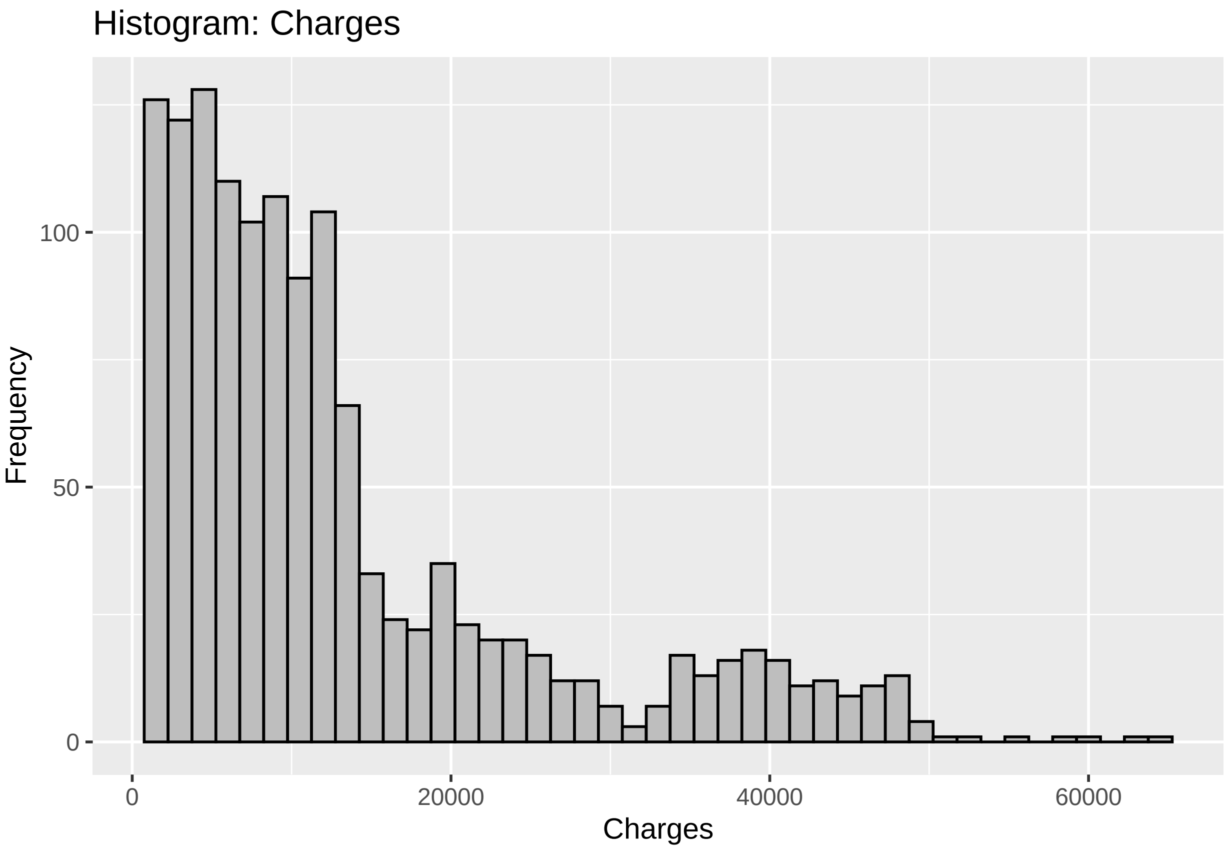
```
children_hist <- ggplot(data = ins_df, aes(x=children)) +
  geom_histogram(color ="black", fill="grey", binwidth = .9) +
  labs(title = "Histogram: Children", x = "Number of Children", y = "Frequency")
children_hist
```



```r
bmi_hist <- ggplot(data = ins_df, aes(x=bmi)) +
  geom_histogram(color ="black", fill="grey", binwidth = 8) +
  labs(title = "Histogram: BMI", x = "Body Mass Index", y = "Frequency")
bmi_hist
```



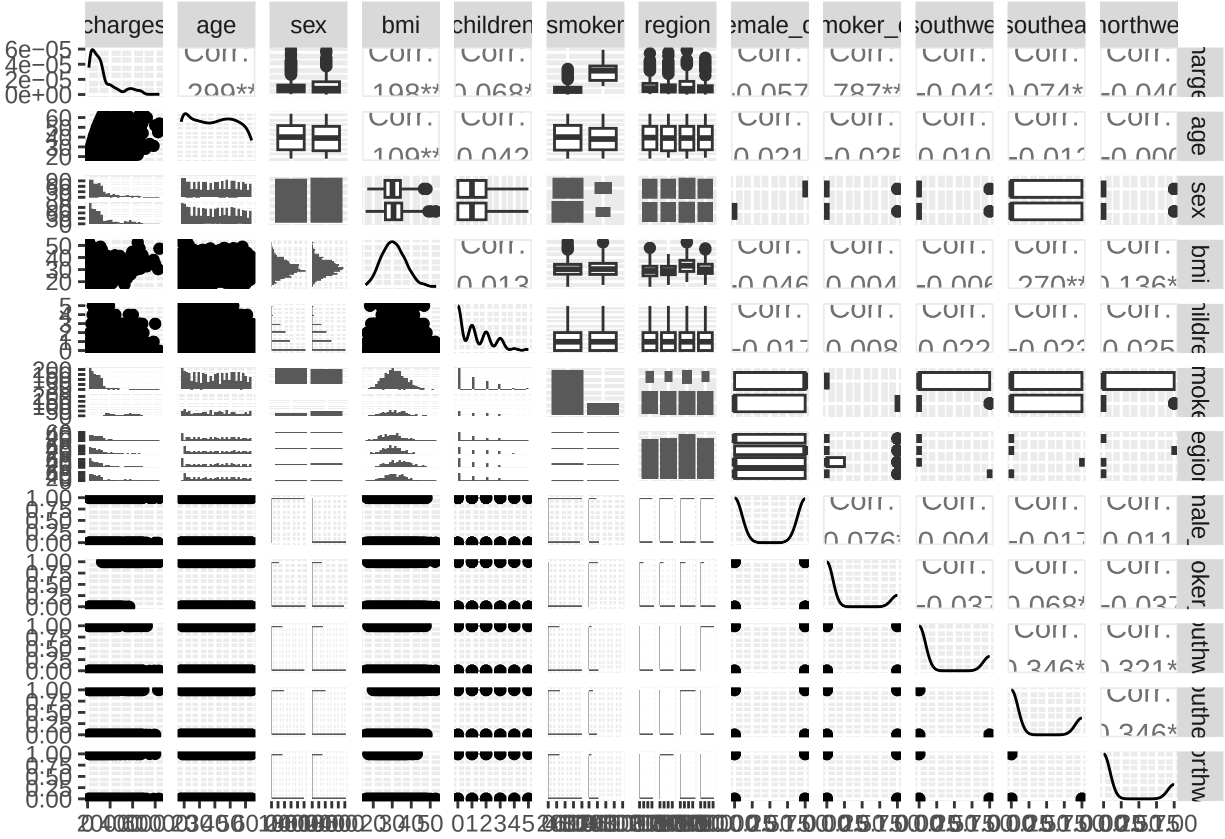
```
charges_hist <- ggplot(data = ins_df, aes(x=charges)) +
  geom_histogram(color ="black", fill="grey", binwidth = 1500) +
  labs(title = "Histogram: Charges", x = "Charges", y = "Frequency")
charges_hist
```



Correlation of Variables and Linear Regression Models

```r
##What variables have the strongest correlation
ggpairs(ins_df[, c(7, 1:6, 8:12)])
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



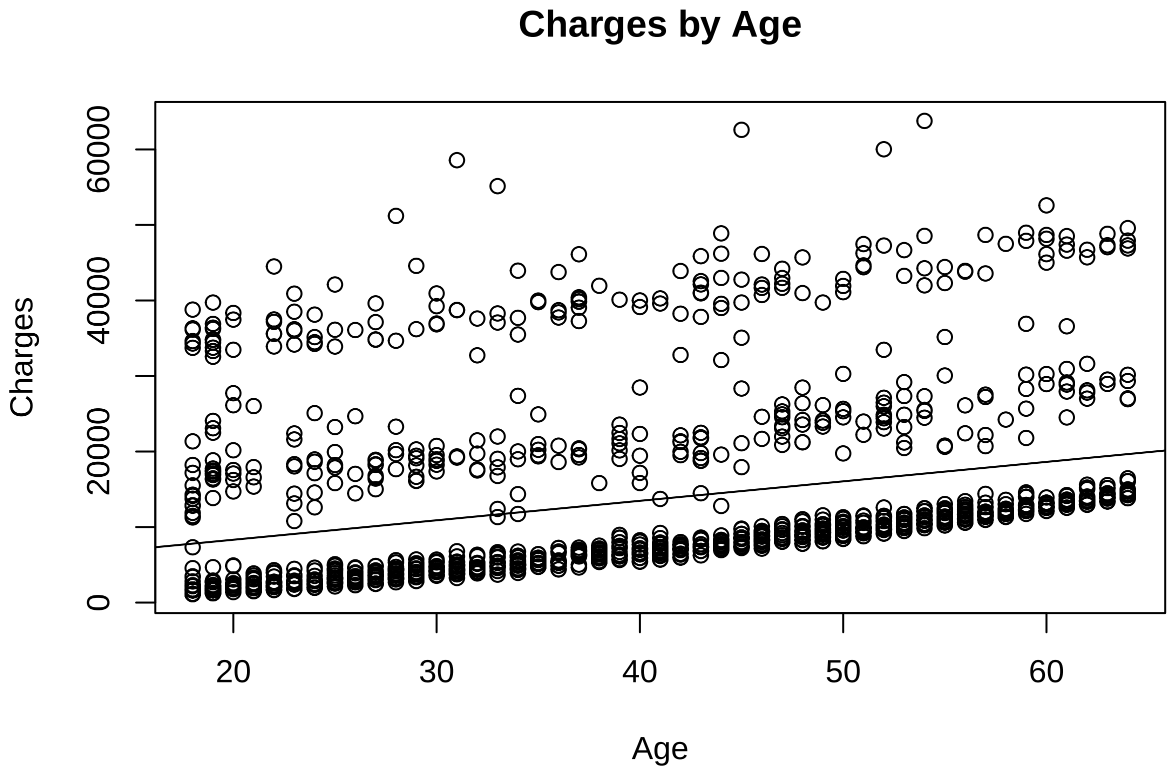
```
##multi-linear regression of all numeric values with dummies included
lm_model <- lm(charges ~ age + bmi + children + female_dt + smoker_dt +
    reg_southwest_dt + reg_southeast_dt + reg_northwest_dt, data = ins_df)
summary(lm_model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + female_dt + smoker_dt +
##     reg_southwest_dt + reg_southeast_dt + reg_northwest_dt, data = ins_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -12069.9      999.6 -12.074  < 2e-16 ***
## age                 256.9       11.9  21.587  < 2e-16 ***
## bmi                 339.2       28.6  11.860  < 2e-16 ***
## children            475.5      137.8   3.451 0.000577 ***
## female_dt           131.3      332.9   0.394 0.693348
## smoker_dt         23848.5      413.1  57.723  < 2e-16 ***
## reg_southwest_dt   -960.0      477.9  -2.009 0.044765 *
## reg_southeast_dt  -1035.0      478.7  -2.162 0.030782 *
## reg_northwest_dt   -353.0      476.3  -0.741 0.458769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

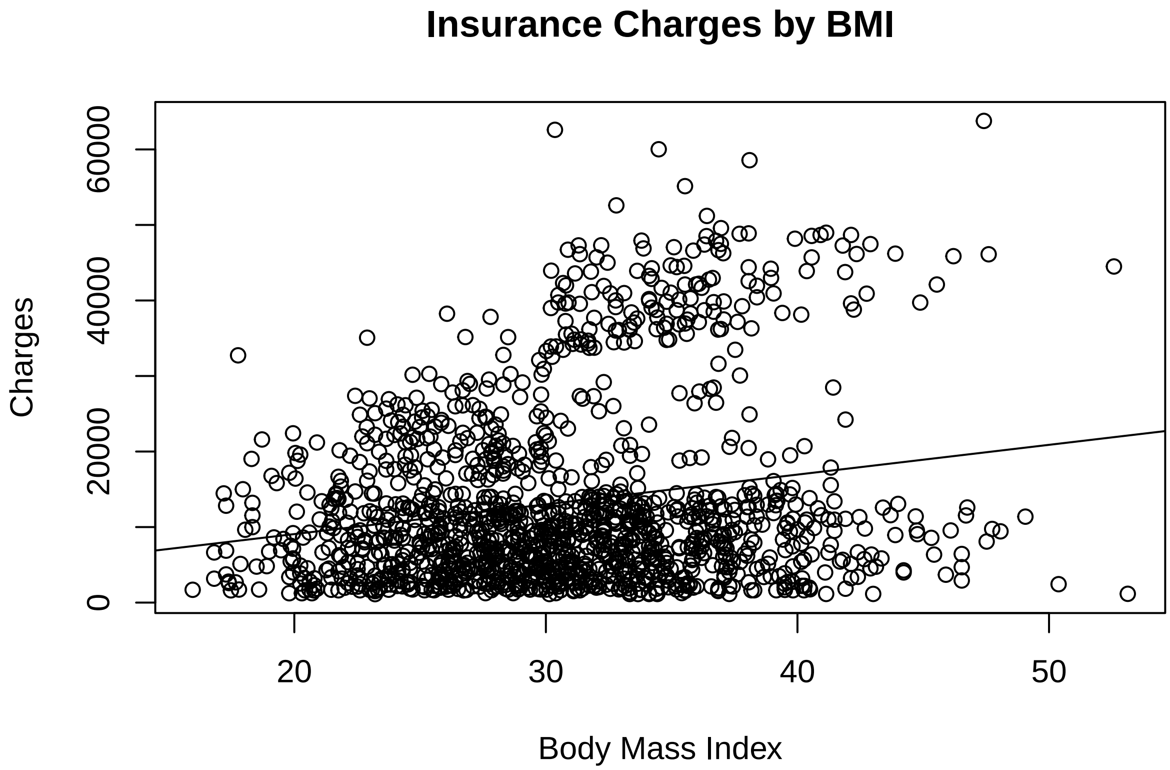
```
##remove insignificant variables and removed the region
##as only 2 variables were significant and should not be used for analysis
lm_model_sign <- lm(charges ~ age + bmi+ children + smoker_dt,
                  data = ins_df)
summary(lm_model_sign)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker_dt, data = ins_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
## age            257.85      11.90  21.675  < 2e-16 ***
## bmi            321.85      27.38  11.756  < 2e-16 ***
## children       473.50     137.79   3.436 0.000608 ***
## smoker_dt    23811.40     411.22  57.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

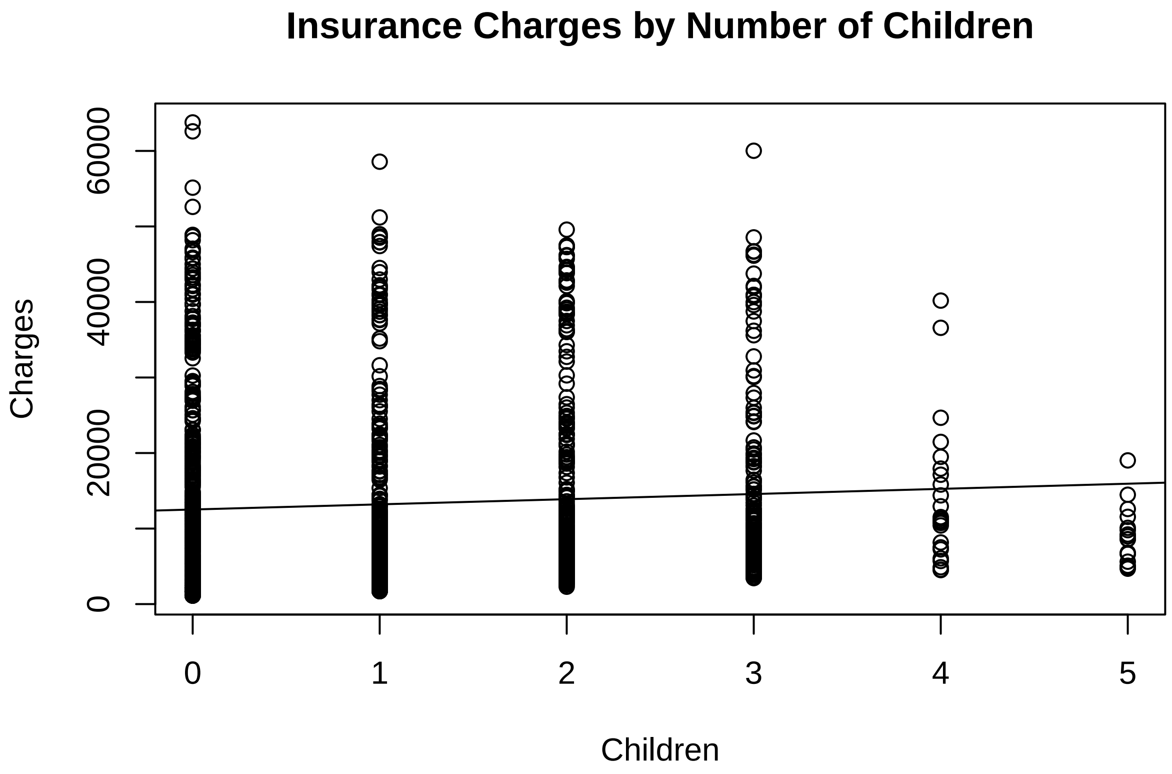
```
lm_model_charge_age <-  lm(charges ~ age, data = ins_df)
plot(ins_df$age,ins_df$charges,
     main = "Charges by Age",
     xlab = "Age",
     ylab = "Charges")
abline(lm_model_charge_age)
```



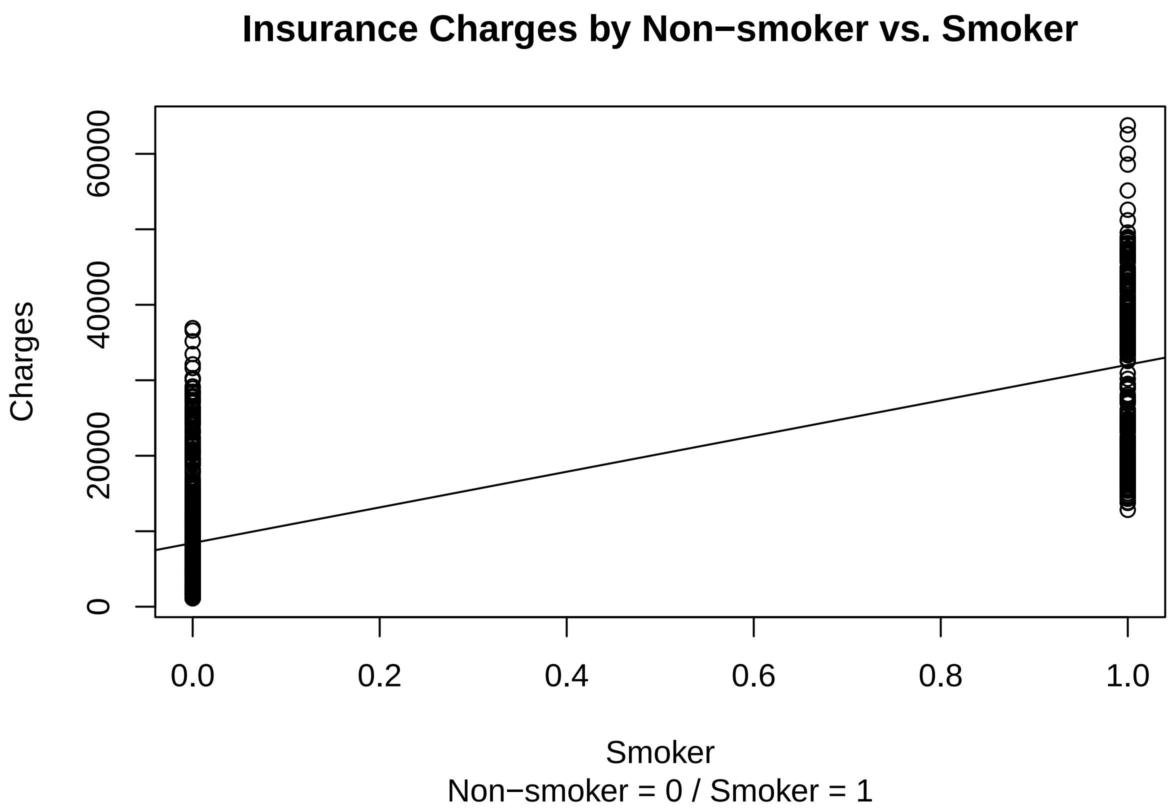
```
lm_model_charge_bmi <-  lm(charges ~ bmi, data = ins_df)
plot(ins_df$bmi,ins_df$charges,
     main = "Insurance Charges by BMI",
     xlab = "Body Mass Index",
     ylab = "Charges")
abline(lm_model_charge_bmi)
```



```r
lm_model_charge_children <- lm(charges ~ children, data = ins_df)
plot(ins_df$children,ins_df$charges,
     main = "Insurance Charges by Number of Children",
     xlab = "Children",
     ylab = "Charges")
abline(lm_model_charge_children)
```



```r
lm_model_charge_smoker <- lm(charges ~ smoker_dt, data = ins_df)
plot(ins_df$smoker_dt,ins_df$charges,
     main = "Insurance Charges by Non-smoker vs. Smoker",
     sub = "Non-smoker = 0 / Smoker = 1",
     xlab = "Smoker",
     ylab = "Charges")
abline(lm_model_charge_smoker)
```



```
#Adjusted R-Squared: 0.7489

#determine which variable has the most individual impact on charges
lm_model_no_age <- lm(charges ~ bmi+ children + smoker_dt, data = ins_df)
summary(lm_model_no_age)
```

```
##
## Call:
## lm(formula = charges ~ bmi + children + smoker_dt, data = ins_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15307.4  -4629.1   -932.1   3744.2  31342.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4064.47    1006.62  -4.038  5.7e-05 ***
## bmi           386.51      31.64  12.217  < 2e-16 ***
## children      597.55     160.05   3.734 0.000197 ***
## smoker_dt   23580.37     477.88  49.343  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7054 on 1334 degrees of freedom
## Multiple R-squared:  0.6615, Adjusted R-squared:  0.6607
## F-statistic: 868.9 on 3 and 1334 DF,  p-value: < 2.2e-16
```

```
#Adjusted R-Squared: 0.6607
```

```r
lm_model_no_bmi <- lm(charges ~ age+ children + smoker_dt, data = ins_df)
summary(lm_model_no_bmi)
```

```
##
## Call:
## lm(formula = charges ~ age + children + smoker_dt, data = ins_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15547.3  -1941.4  -1319.1   -425.3  29313.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2851.99     543.78  -5.245 1.82e-07 ***
## age           273.09      12.42  21.990  < 2e-16 ***
## children      486.65     144.70   3.363 0.000792 ***
## smoker_dt   23842.60     431.84  55.212  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6372 on 1334 degrees of freedom
## Multiple R-squared:  0.7237, Adjusted R-squared:  0.7231
## F-statistic:  1165 on 3 and 1334 DF,  p-value: < 2.2e-16
```

```r
#Adjusted R-Squared: 0.7231

lm_model_no_children <- lm(charges ~ bmi+ age + smoker_dt, data = ins_df)
summary(lm_model_no_children)
```

```
##
## Call:
## lm(formula = charges ~ bmi + age + smoker_dt, data = ins_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12415.4  -2970.9   -980.5   1480.0  28971.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11676.83     937.57  -12.45   <2e-16 ***
## bmi            322.62      27.49   11.74   <2e-16 ***
## age            259.55      11.93   21.75   <2e-16 ***
## smoker_dt    23823.68     412.87   57.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6092 on 1334 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7469
## F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

```r
#Adjusted R-Squared: 0.7469

lm_model_no_smoker <- lm(charges ~ age+bmi+ children, data = ins_df)
summary(lm_model_no_smoker)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children, data = ins_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13884  -6994  -5092   7125  48627
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6916.24    1757.48  -3.935 8.74e-05 ***
## age           239.99      22.29  10.767  < 2e-16 ***
## bmi           332.08      51.31   6.472 1.35e-10 ***
## children      542.86     258.24   2.102   0.0357 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11370 on 1334 degrees of freedom
## Multiple R-squared:  0.1201, Adjusted R-squared:  0.1181
## F-statistic: 60.69 on 3 and 1334 DF,  p-value: < 2.2e-16
```

```
#Adjusted R-Squared: 0.1181
#Suggests smoking has the greatest individual impact on the model,
##followed by age.
```

Predictive Model and Function

```
##Prediction model of charges based off of significant variables
##Random variable inputs to test prediction
pred_df <- data.frame(age = 18, bmi = 20, children = 0, smoker_dt = 1)
predict(lm_model_sign, pred_df,type = "response")
```

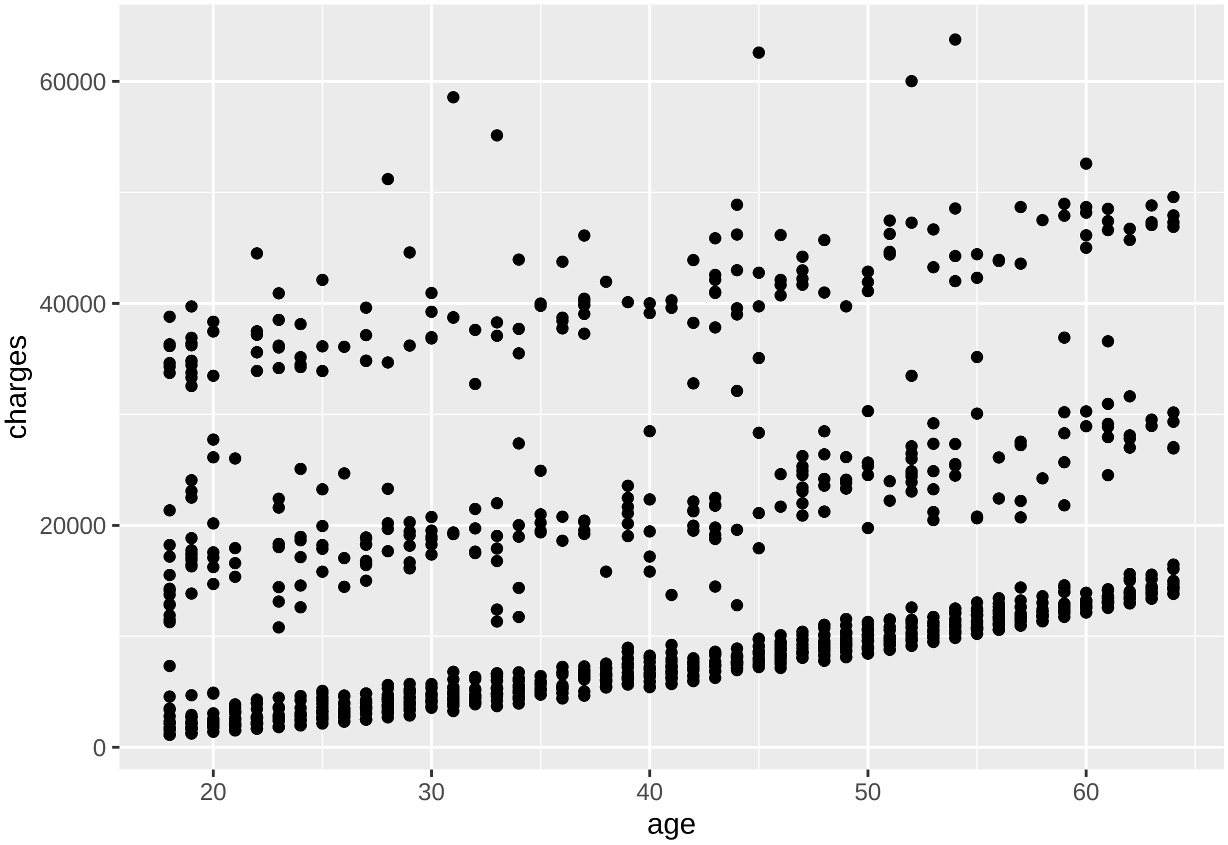
```
##        1
## 22786.95
```

```
##Create a function for various inputs to predict charges based off of individual variables
charges_function <- function(age,bmi,children,smoker) {
  model_charges <- lm(charges ~ age + bmi+ children + smoker_dt,
                      data = ins_df)
  pred_df <- data.frame(age = age, bmi = bmi, children = children,
                        smoker_dt = smoker)
  pred_charges <- predict(model_charges,pred_df, type = "response")
  return(pred_charges)
}
##example data, output is $5,631.92
charges_function(27,32,1,0)
```

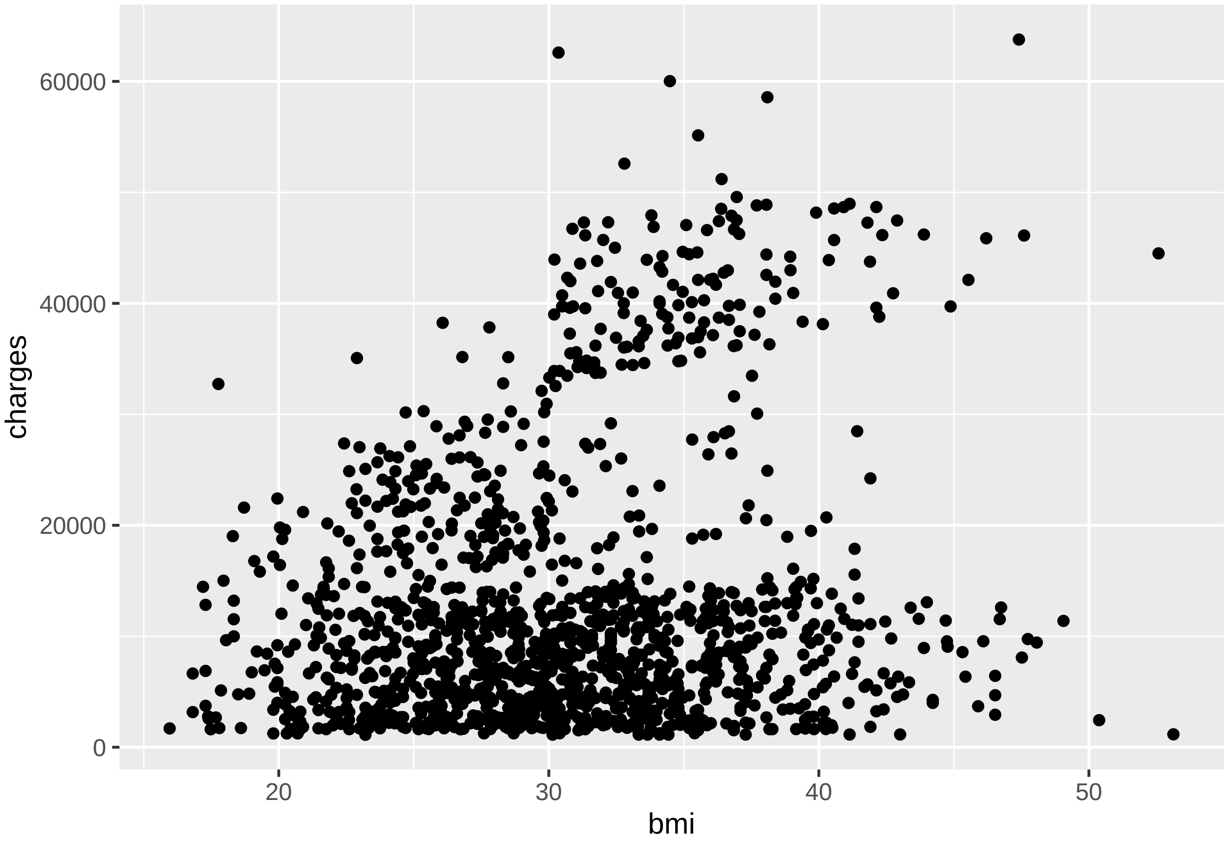
```
##        1
## 5631.915
```

Plots

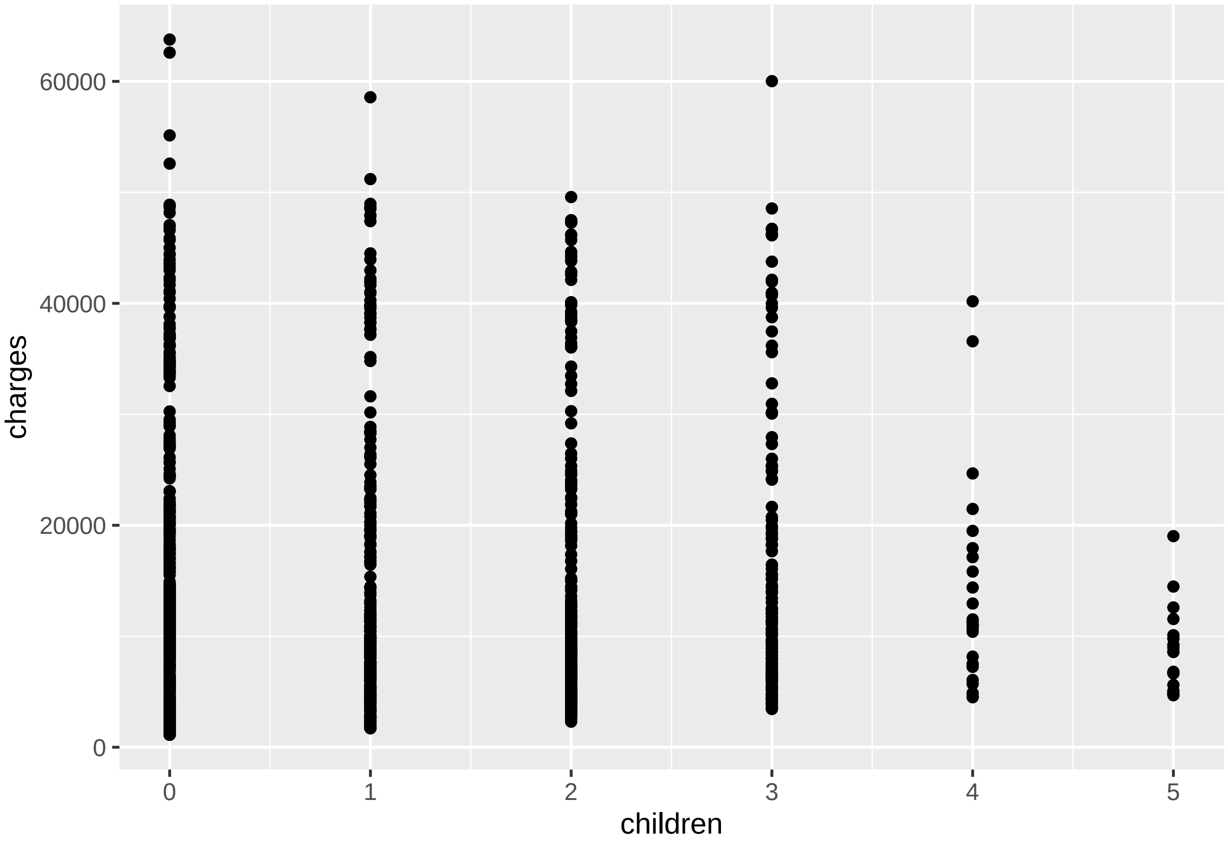
```
##scatter plot some variables for initial analysis
plot_age <- ggplot(ins_df, aes(x=age,y=charges)) + geom_point()
plot_age
```



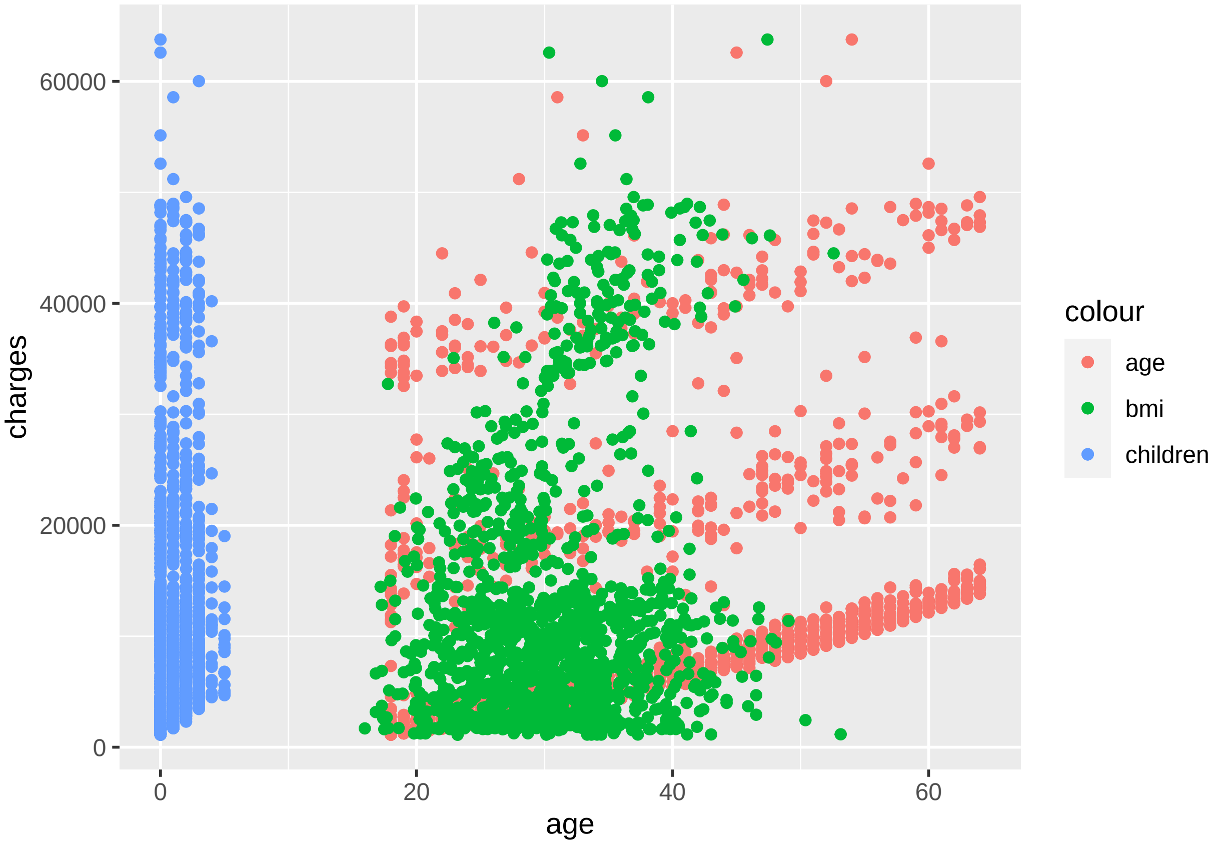
```
plot_bmi <- ggplot(ins_df, aes(x=bmi,y=charges)) + geom_point()
plot_bmi
```



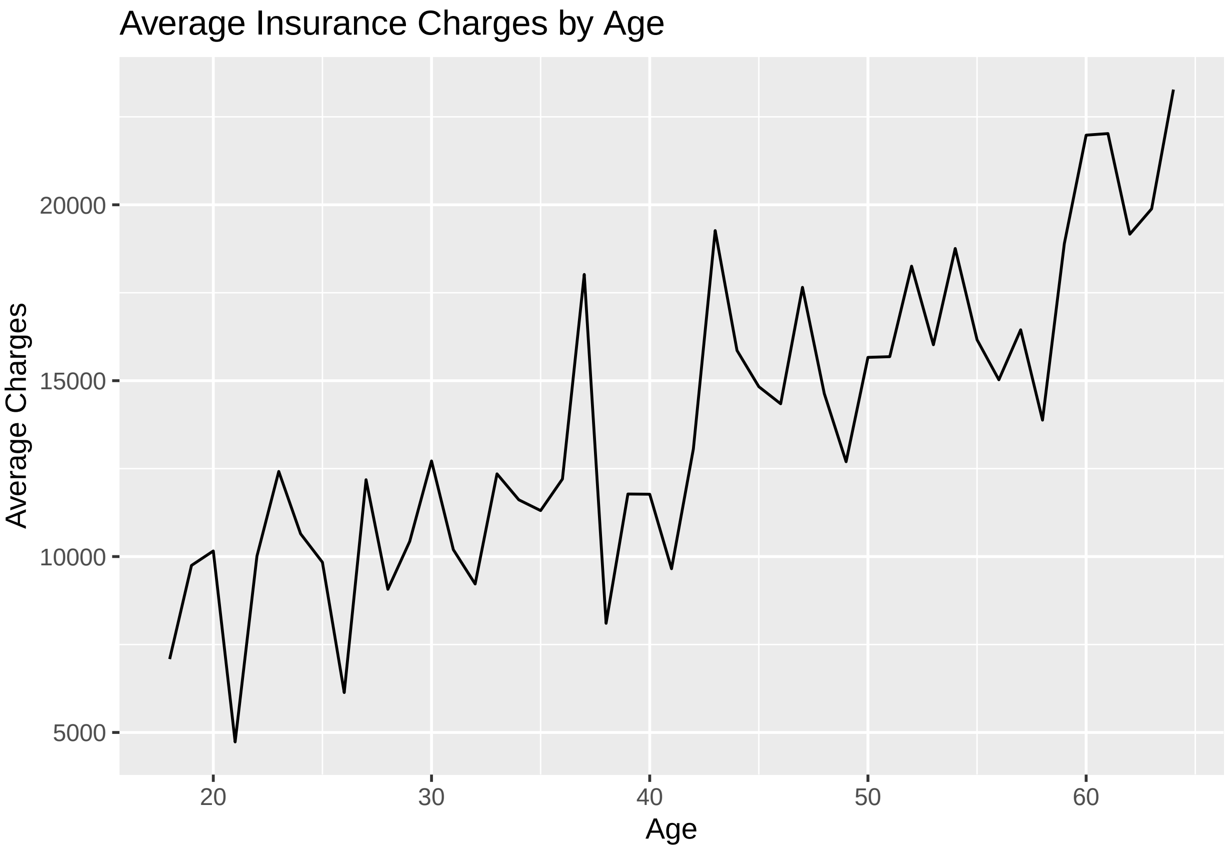
```
plot_children <- ggplot(ins_df, aes(x=children,y=charges)) + geom_point()
plot_children
```



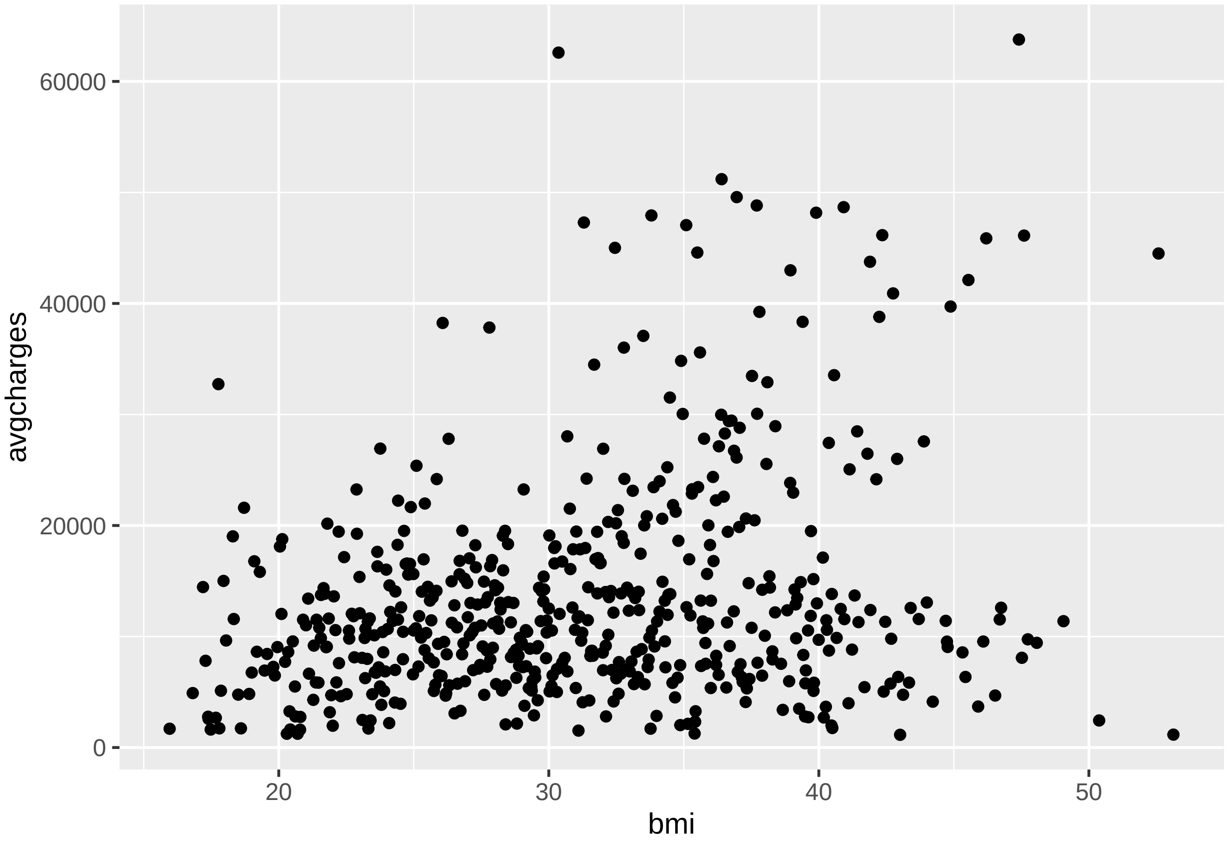
```r
#Visualization of all variables in same plot deemed not useful
plot_all <- ggplot(ins_df,aes(y=charges)) +
  geom_point(aes(x = age, color = "age")) +
  geom_point(aes(x = bmi, color = "bmi")) +
  geom_point(aes(x = children, color = "children"))
plot_all
```



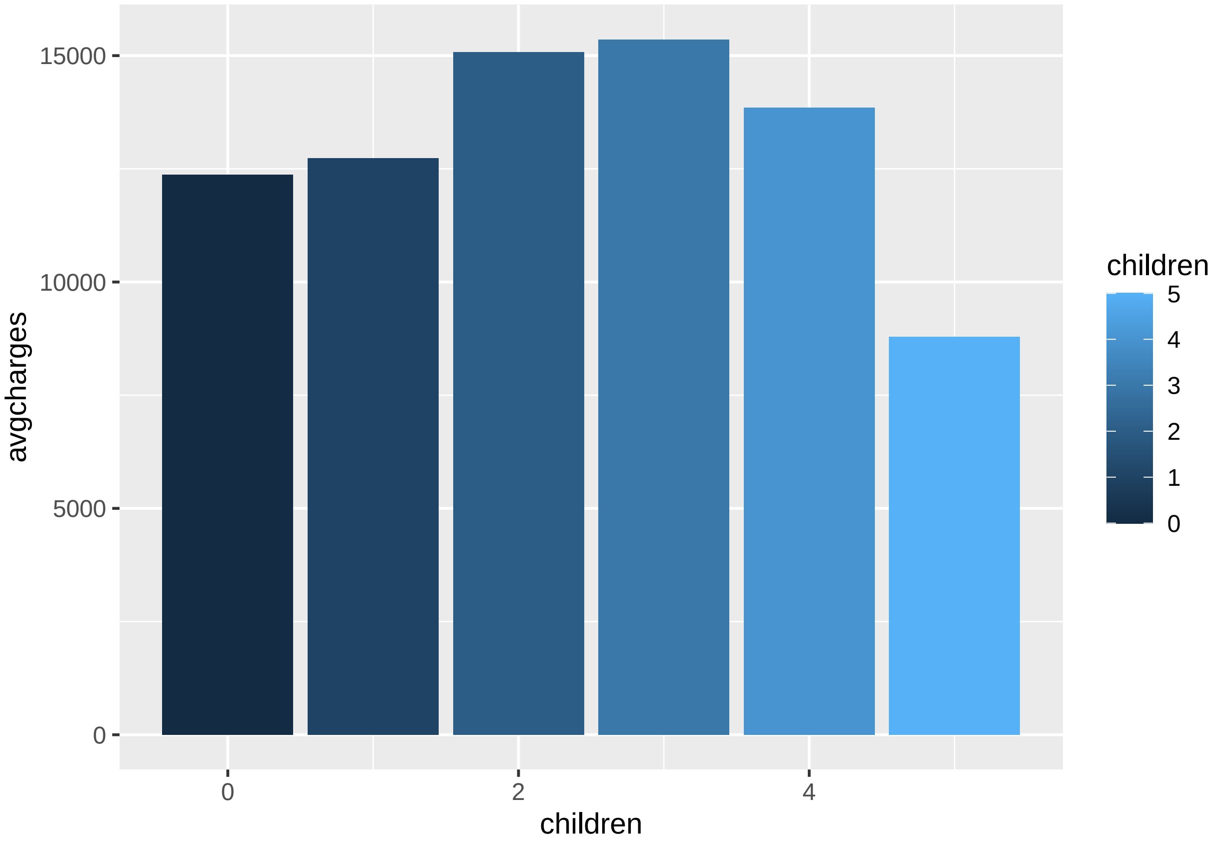
```
##line plot of average charges by age
avg_charges_by_age <- sqldf("select age, avg(charges) as avgcharges
                from ins_df group by age order by age desc")
plot_avg_charges_by_age <- ggplot(avg_charges_by_age, aes(y=avgcharges)) +
  geom_line(aes(x = age)) +
  labs(title = "Average Insurance Charges by Age",
       x = "Age", y = "Average Charges")
plot_avg_charges_by_age
```



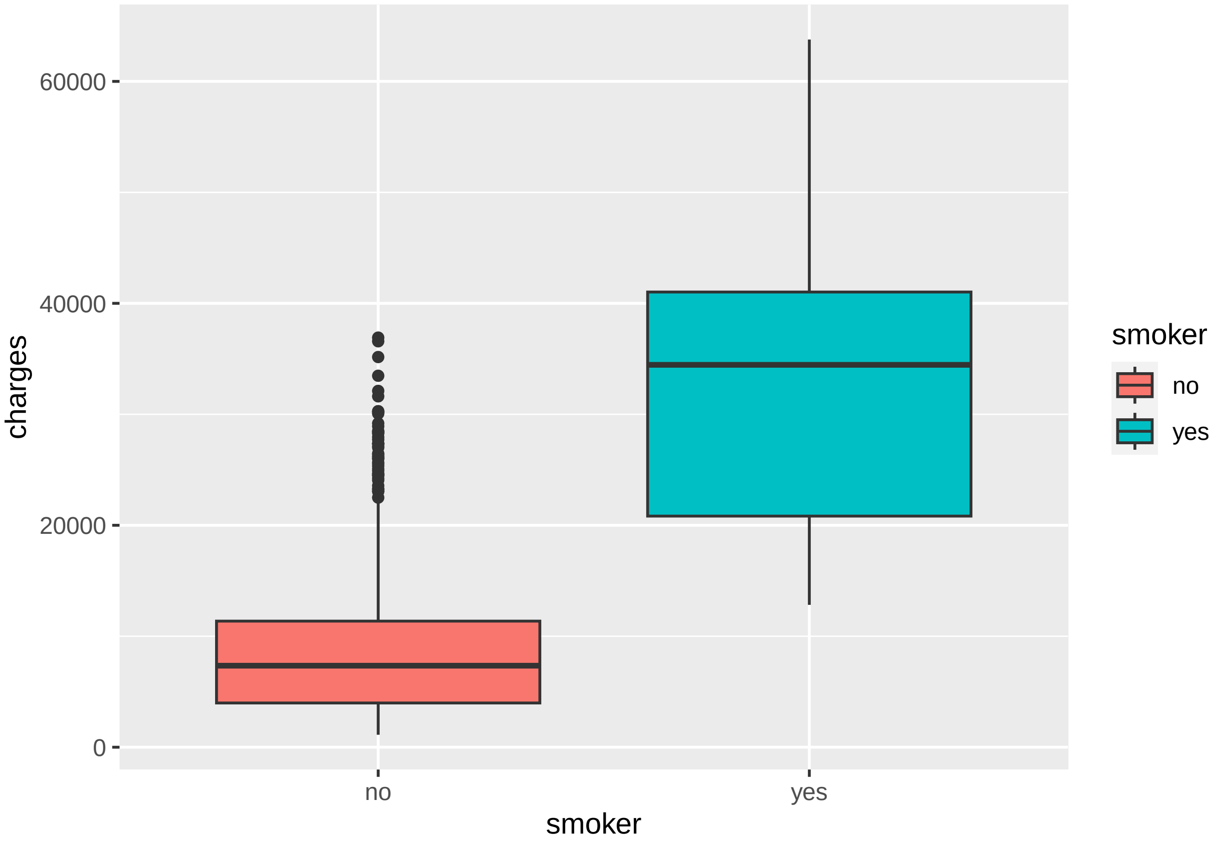
```
##line* plot of average charges by bmi

##Replaced line with scatter, added color shift as possible solution -NB
avg_charges_by_bmi <- sqldf("select bmi, avg(charges) as avgcharges
                  from ins_df group by bmi order by bmi desc")
plot_avg_charges_by_bmi <- ggplot(avg_charges_by_bmi, aes(y=avgcharges)) +
  geom_point(aes(x = bmi))
plot_avg_charges_by_bmi
```



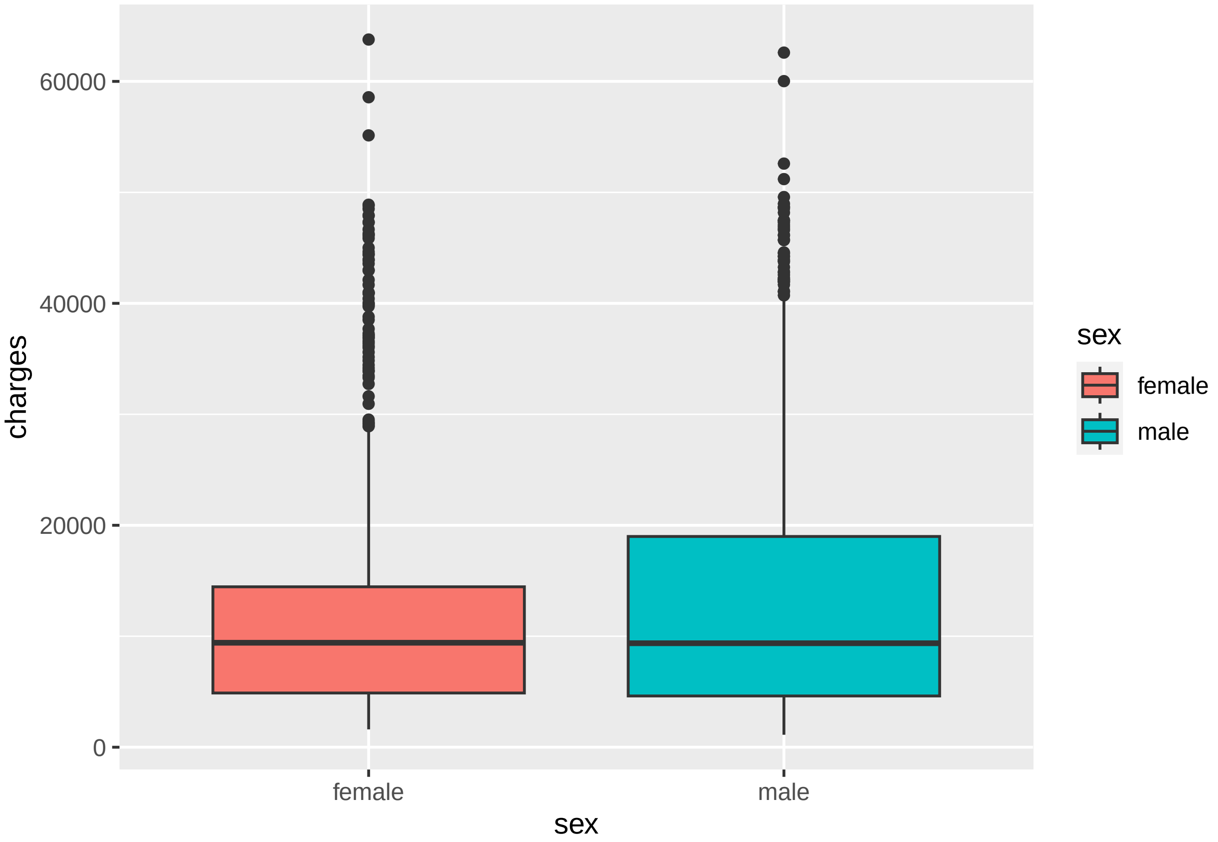
```
#column plot of average charges by children
##Suggests that having 2-4 children increases charge, but overall number of
##children does not greatly impact charges (based on findings from
##regression models)
avg_charges_by_children <- sqldf("select children, avg(charges)
      as avgcharges from ins_df group by children order by children desc")
plot_avg_charges_by_children <- ggplot(avg_charges_by_children,
              aes(y=avgcharges,fill=children)) +geom_col(aes(x = children))
plot_avg_charges_by_children
```



```
##boxplot of charges for smokers vs non-smokers
##Shows cluster of outliers for non-smokers, suggesting another factor
##is determining charges for those patients.
box_smoker_charges <- ggplot(ins_df,aes(x=smoker,y=charges,fill=smoker)) +
  geom_boxplot()
box_smoker_charges
```



```
#boxplot of charges for females vs. males
#suggests the median charge is about the same, but that female patients
##have more outliers, while male patients have a wider range of average charges
#earlier model suggests this is not significant overall
box_sex_charges <- ggplot(ins_df,aes(x=sex,y=charges,fill=sex)) +
  geom_boxplot()
box_sex_charges
```



Shiny App

```
#Shiny app (interactive predictive function)
sideBar1 <- sidebarPanel(
  numericInput("age","Age of patient",value=0,min = 0,max = 130,step=1),width=8)
sideBar2 <- sidebarPanel(
  numericInput("bmi","BMI of patient",value=12,min = 12,max = 210,step=1),width=8)
sideBar3 <- sidebarPanel(
  numericInput("children",
               "Number of children",value=0,min = 0,max = 20,step=1),width=8)
sideBar4<- sidebarPanel(
  radioButtons("smoker_dt","Is Patient a Smoker?",choices=c("Yes","No"),selected="Yes"),width=8)
ui <- fluidPage(
  theme=shinytheme("readable"),
  titlePanel("Charges"),
  fluidRow(sideBar1),
  fluidRow(sideBar2),
  fluidRow(sideBar3),
  fluidRow(sideBar4),
  fluidRow(
    column(5,
           mainPanel(actionButton("button", "Calculate"),wellPanel(textOutput("Charges"))))))

server <- function(input,output,session){
  insurance <- read_csv("insurance.csv")
  ins_df <- insurance
  ins_df <- cbind(ins_df,female_dt,smoker_dt,reg_southwest_dt,
                  reg_southeast_dt,reg_northwest_dt)
```

```r
  press <- eventReactive(input$button,{
    charges_function <- function(age,bmi,children,smoker){
      model_charges <- lm(charges ~ age + bmi+ children + smoker_dt,
                          data = ins_df)
      pred_df <- data.frame(age = age, bmi = bmi, children = children,
                        smoker_dt=as.numeric(smoker_dt))
      pred_charges <- predict(model_charges,pred_df, type = "response")
      return(mean(pred_charges))
    }
    charges_function(input$age,input$bmi,input$children,input$smoker_dt)})
  output$Charges <- renderText({press()})
}
shinyApp(ui, server)
```