Untitled

Zhiwei Lin

2023-01-24

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## 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(skimr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(GGally)
```

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

```
library(ggpmisc)
```

```
## Loading required package: ggpp
##
## Attaching package: 'ggpp'
##
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

load dataset

```
data<-read.csv("/Users/zhiweilin/Downloads/insurance.csv", header =T, na.string=c("","NA"))
```

```
head(data)
```

```
##   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
```

```
str(data)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

View the first 6 rows of the dataset and the class of each variable in the dataset.

```
data <- mutate_at(data, vars(sex,smoker,region), as.factor)
```

The variables 'sex', 'smoker', and 'region' should be converted from character variables to factor variables as they have a limited number of levels.

```
data<-distinct(data)
```

Remove duplicate rows based on all columns

Data Summarization

```
summary(data)
```

```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1063
##  1st Qu.:27.00   male  :675   1st Qu.:26.29   1st Qu.:0.000   yes: 274
##  Median :39.00                Median :30.40   Median :1.000
##  Mean   :39.22                Mean   :30.66   Mean   :1.096
##  3rd Qu.:51.00                3rd Qu.:34.70   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:324   1st Qu.: 4746
##  southeast:364   Median : 9386
##  southwest:325   Mean   :13279
##                  3rd Qu.:16658
##                  Max.   :63770
```

```
skim_without_charts(data) # another summary function
```

Table 1: Data summary

Name	data
Number of rows	1337
Number of columns	7
Column type frequency:	
factor	3
numeric	4
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
sex	0	1	FALSE	2	mal: 675, fem: 662
smoker	0	1	FALSE	2	no: 1063, yes: 274
region	0	1	FALSE	4	sou: 364, sou: 325, nor: 324, nor: 324

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1	39.22	14.04	18.00	27.00	39.00	51.00	64.00
bmi	0	1	30.66	6.10	15.96	26.29	30.40	34.70	53.13
children	0	1	1.10	1.21	0.00	0.00	1.00	2.00	5.00
charges	0	1	13279.12	12110.36	1121.87	4746.34	9386.16	16657.72	63770.43

The summary function shows that there is no missing values in the dataset.

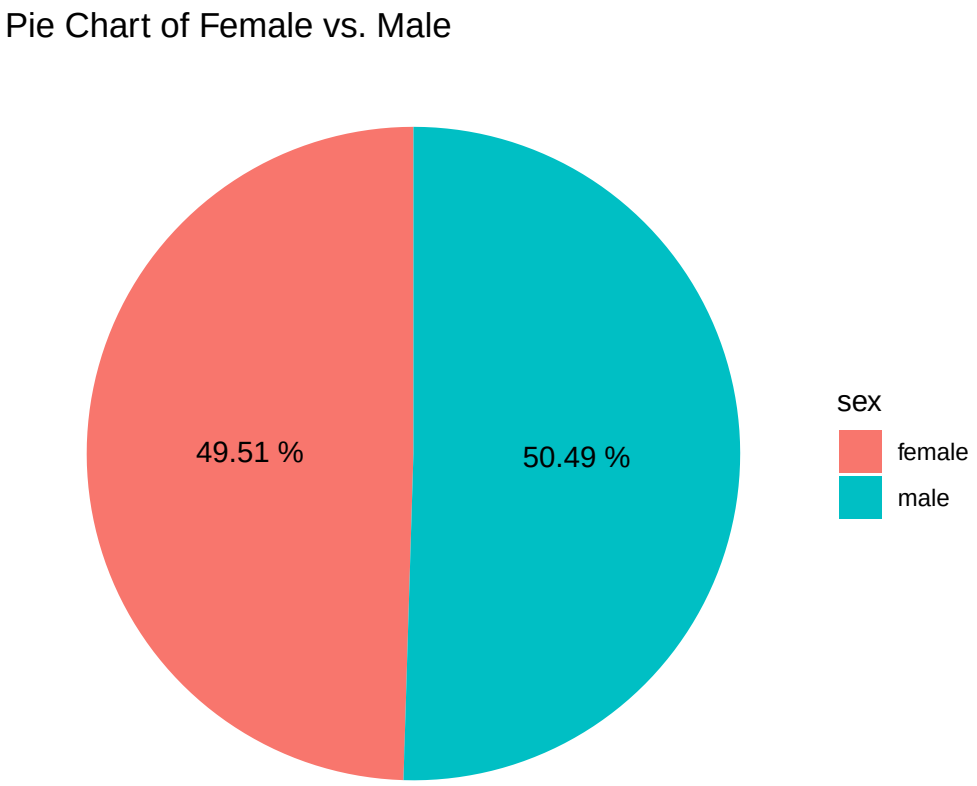
Data Visualization

```
df_sex<-data %>%
  group_by(sex) %>%
  summarise(
    count=n()
  )
df_sex$percentage <- 100*prop.table(df_sex$count)

print(df_sex)
```

```
## # A tibble: 2 x 3
##   sex    count percentage
##   <fct>  <int>      <dbl>
## 1 female   662       49.5
## 2 male     675       50.5
```

```
ggplot(df_sex, aes(x="", y=percentage, fill=sex)) +
  geom_bar(width=1,stat="identity") +
  coord_polar(theta="y", start=0) +
  theme_void()+
  labs(title="Pie Chart of Female vs. Male ", fill="sex")+
   geom_text(aes(label = paste(round(percentage,2), "%")),position = position_stack(vjust = 0.5),color =
```



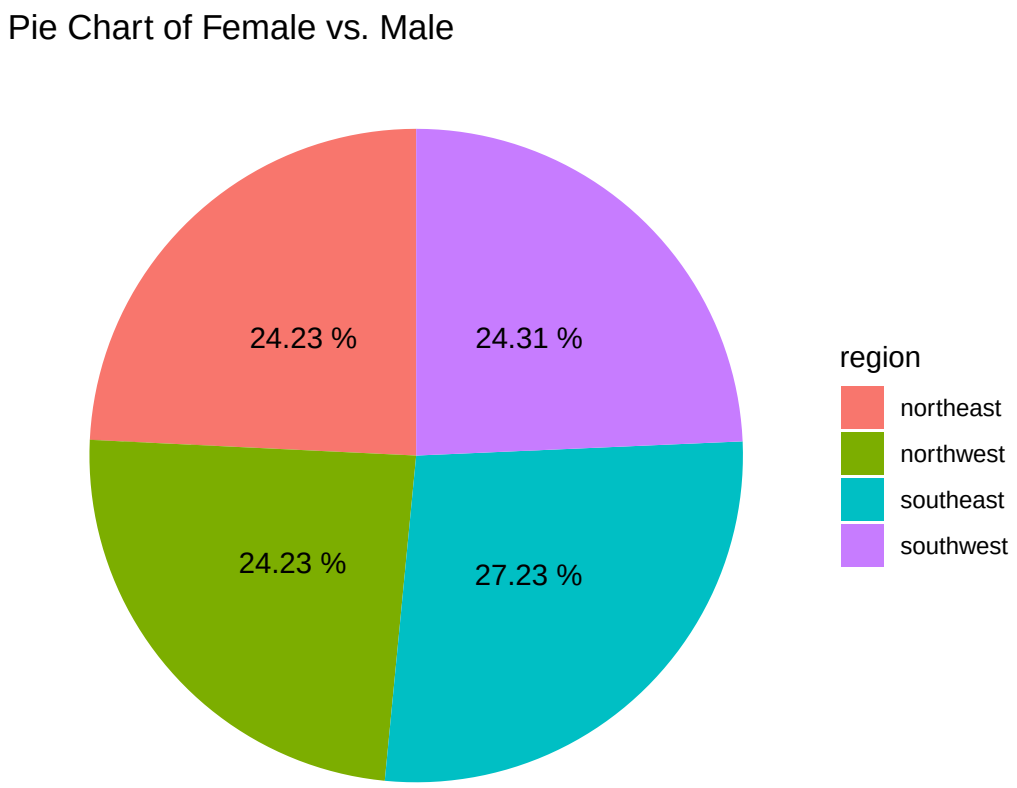
The dataset shows a roughly equal proportion of male and female participants, with 49.51% being female and 50.49% being male, indicating that the data is not biased towards any one gender.

```
df_region<-data %>%
  group_by(region) %>%
  summarise(
    count=n()
  )
df_region$percentage <- 100*prop.table(df_region$count)

print(df_region)
```

```
## # A tibble: 4 x 3
##   region    count percentage
##   <fct>     <int>      <dbl>
## 1 northeast   324       24.2
## 2 northwest   324       24.2
## 3 southeast   364       27.2
## 4 southwest   325       24.3
```

```
ggplot(df_region, aes(x="", y=percentage, fill=region)) +
  geom_bar(width=1,stat="identity") +
  coord_polar(theta="y", start=0) +
  theme_void()+
  labs(title="Pie Chart of Female vs. Male ", fill="region")+
   geom_text(aes(label = paste(round(percentage,2), "%")),position = position_stack(vjust = 0.5),color =
```



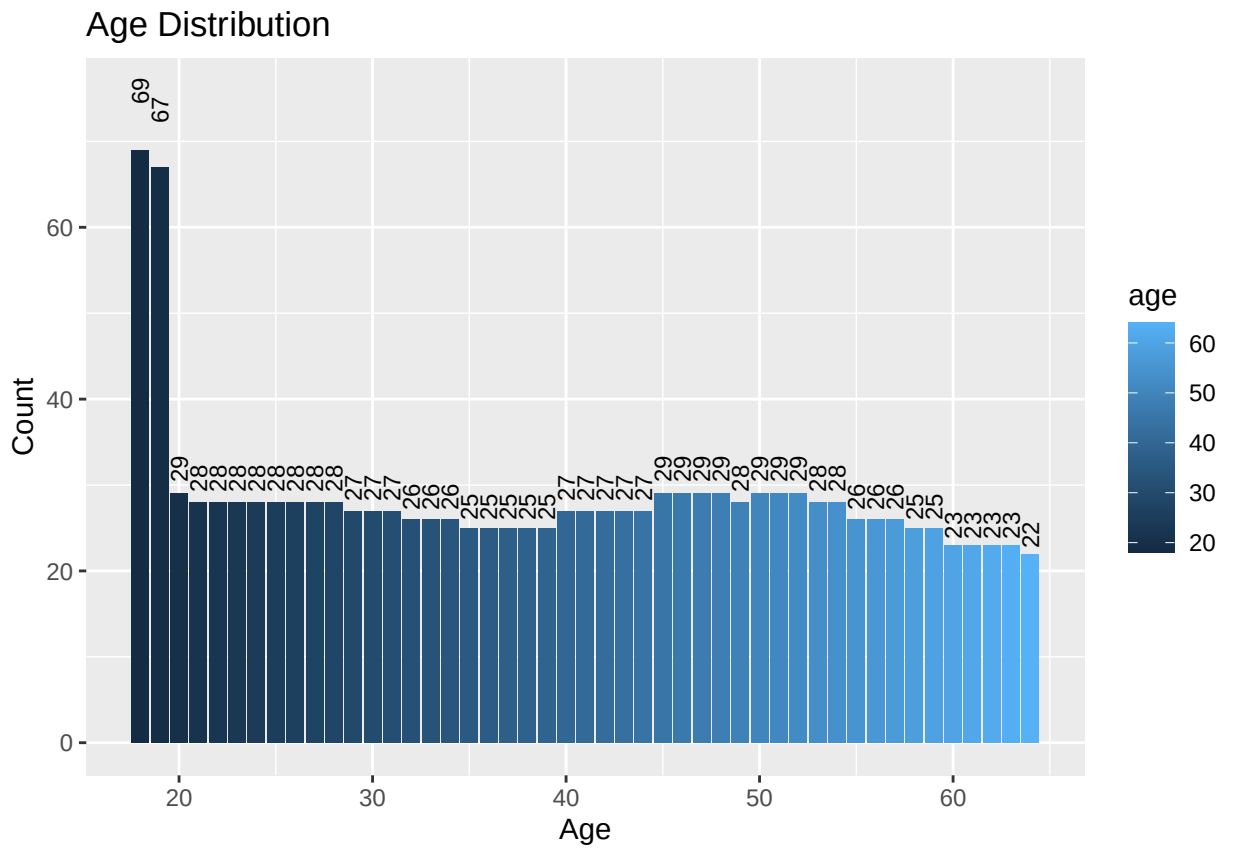
the dataset shows roughly equal proportion of four regions.

```
df_age<-data %>%
  group_by(age) %>%
  summarise(
    number=n()
  )
print(df_age)
```

```
## # A tibble: 47 x 2
##      age number
##    <int>  <int>
##  1    18     69
##  2    19     67
##  3    20     29
##  4    21     28
##  5    22     28
##  6    23     28
##  7    24     28
##  8    25     28
##  9    26     28
## 10    27     28
## # ... with 37 more rows
```

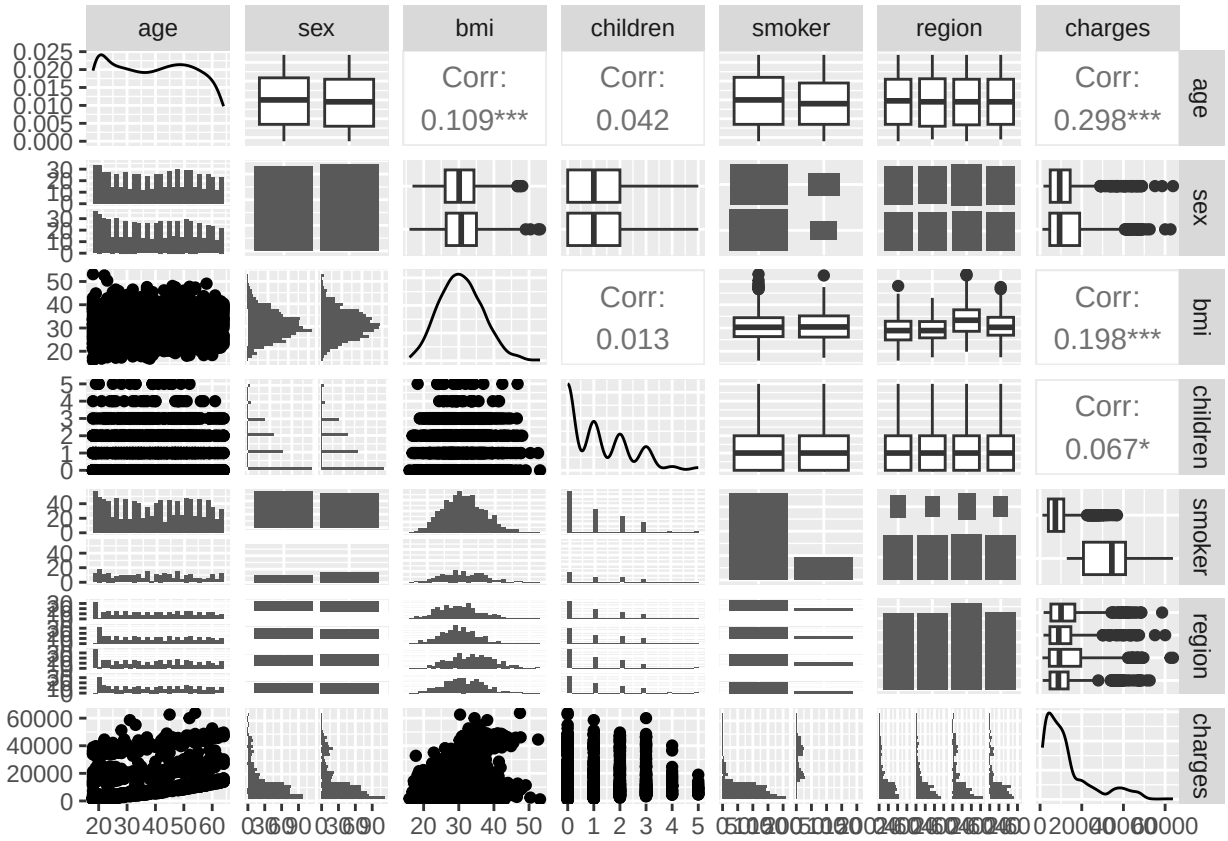
```
ggplot(df_age, aes(x=age,y=number, fill=age)) +
  geom_bar(stat="identity") +
```

```
  geom_text(aes(label = number), position = position_stack(vjust=1.1), angle = 90, size = 3)+
  ggtitle("Age Distribution") +
  xlab("Age") +
  ylab("Count")
```



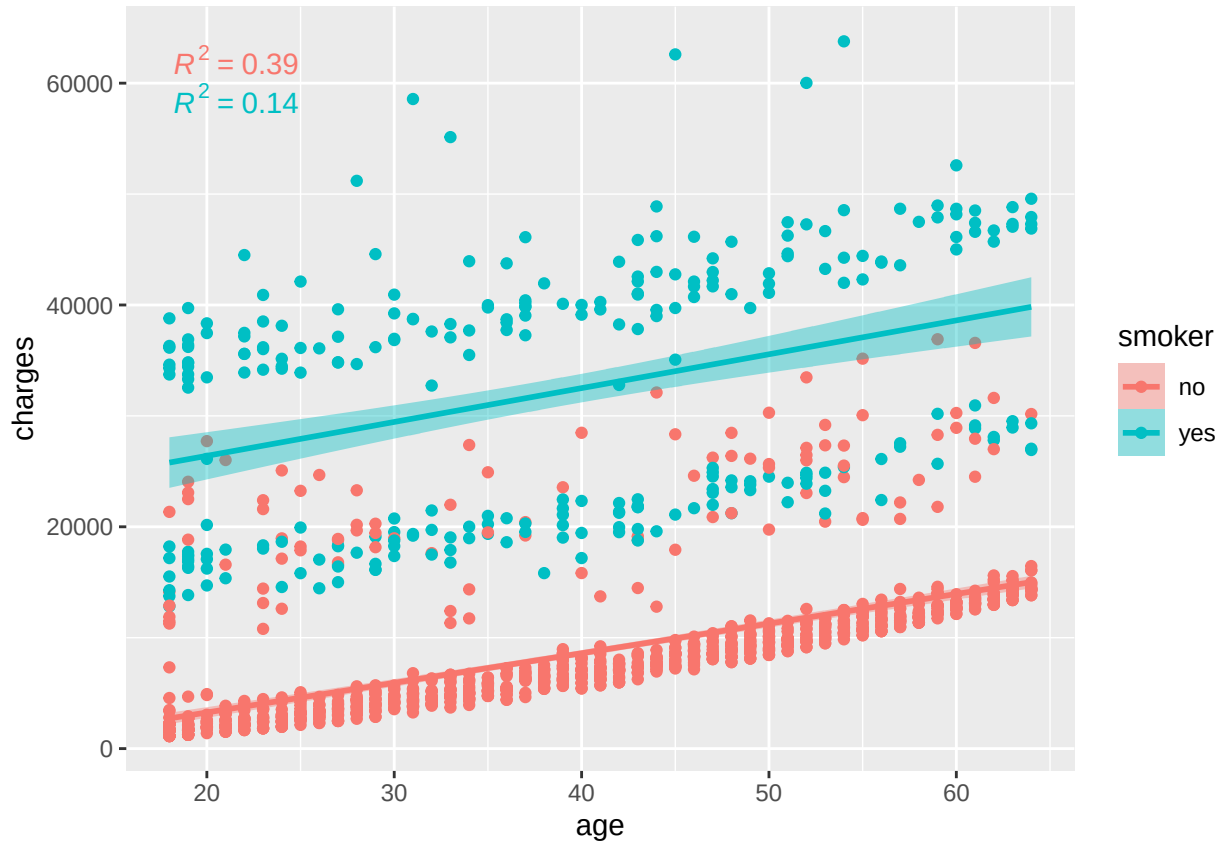
The dataset includes participants aged between 18 and 64, with a majority of around 25 participants per age group. Notably, the age groups of 18 and 19 stand out with more participants, about 65 individuals, than the other age groups.

```
ggpairs(data)
```

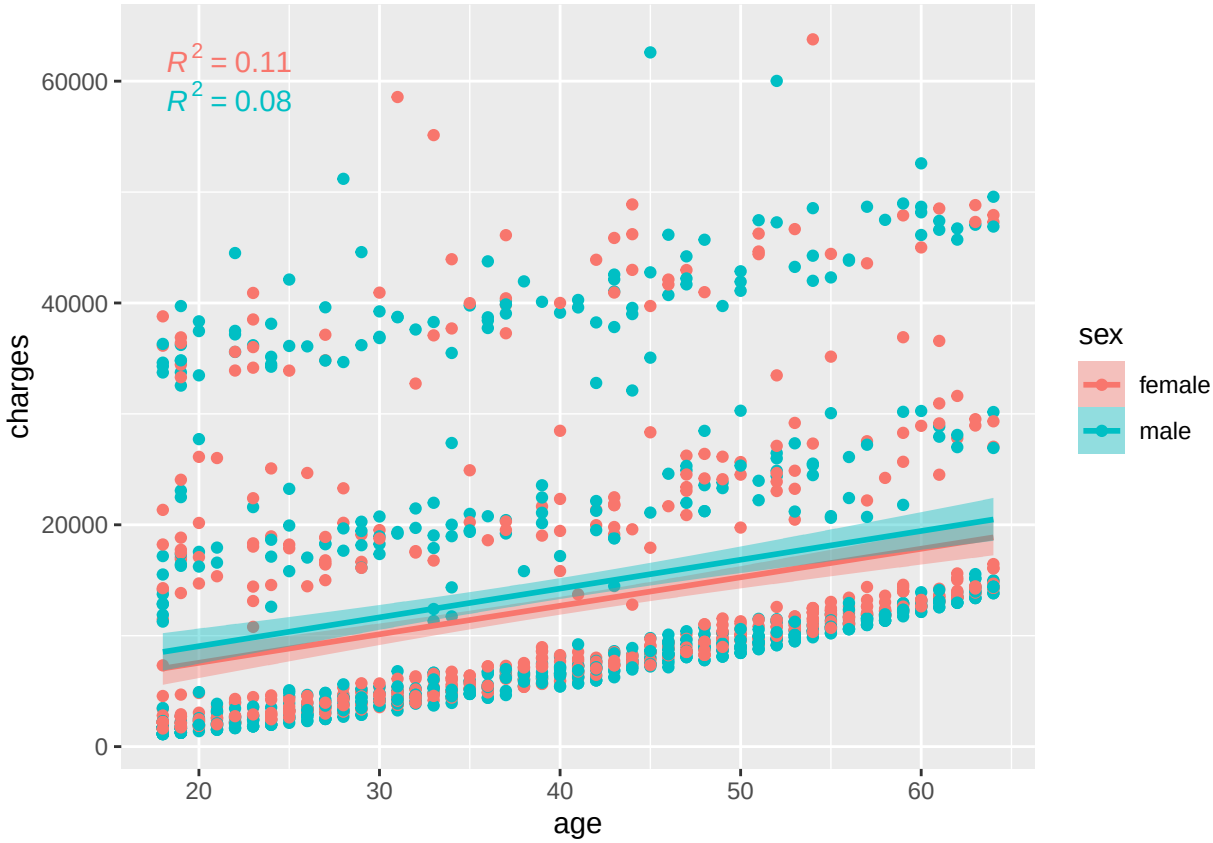


scatterplot matrix with all variables in the dataset

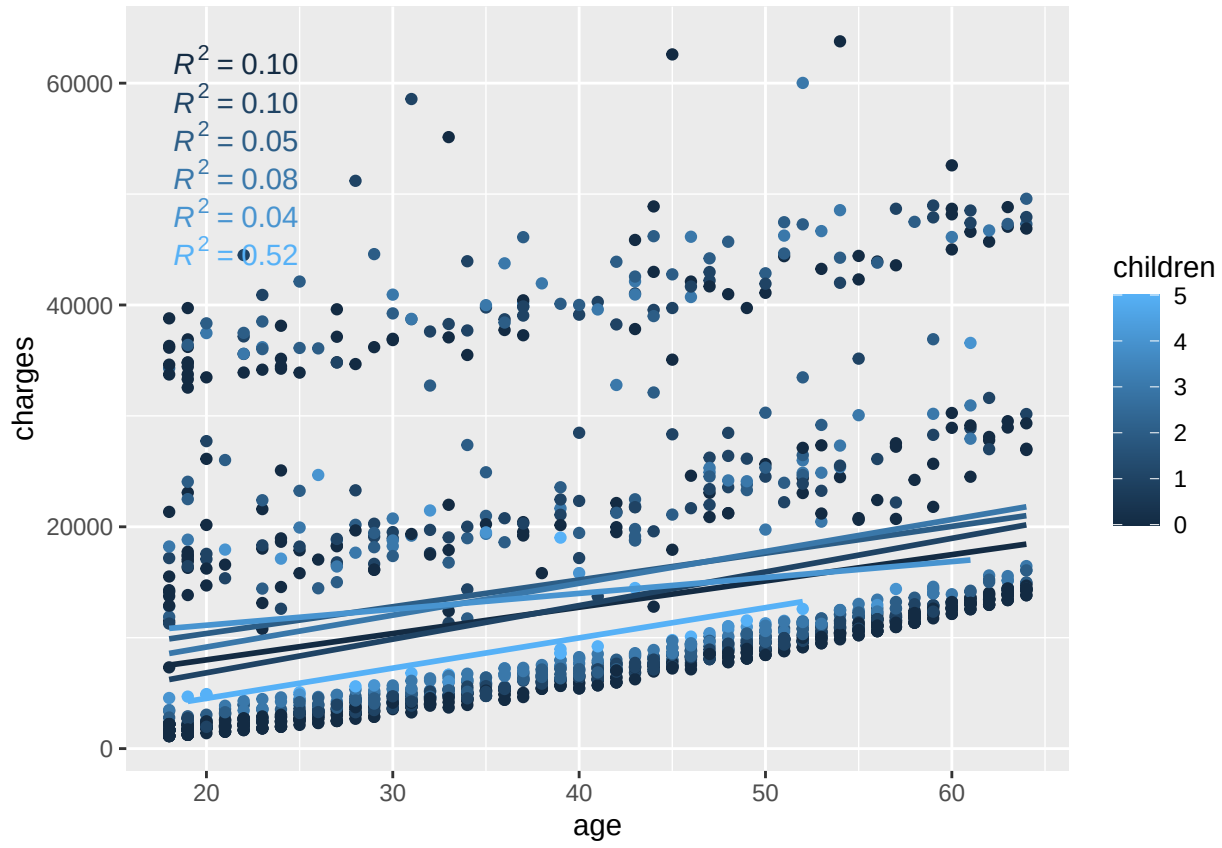
```
ggplot(data,aes(x=age,y=charges,fill=smoker, color=smoker))+geom_point()+ stat_poly_line() + stat_poly_
```



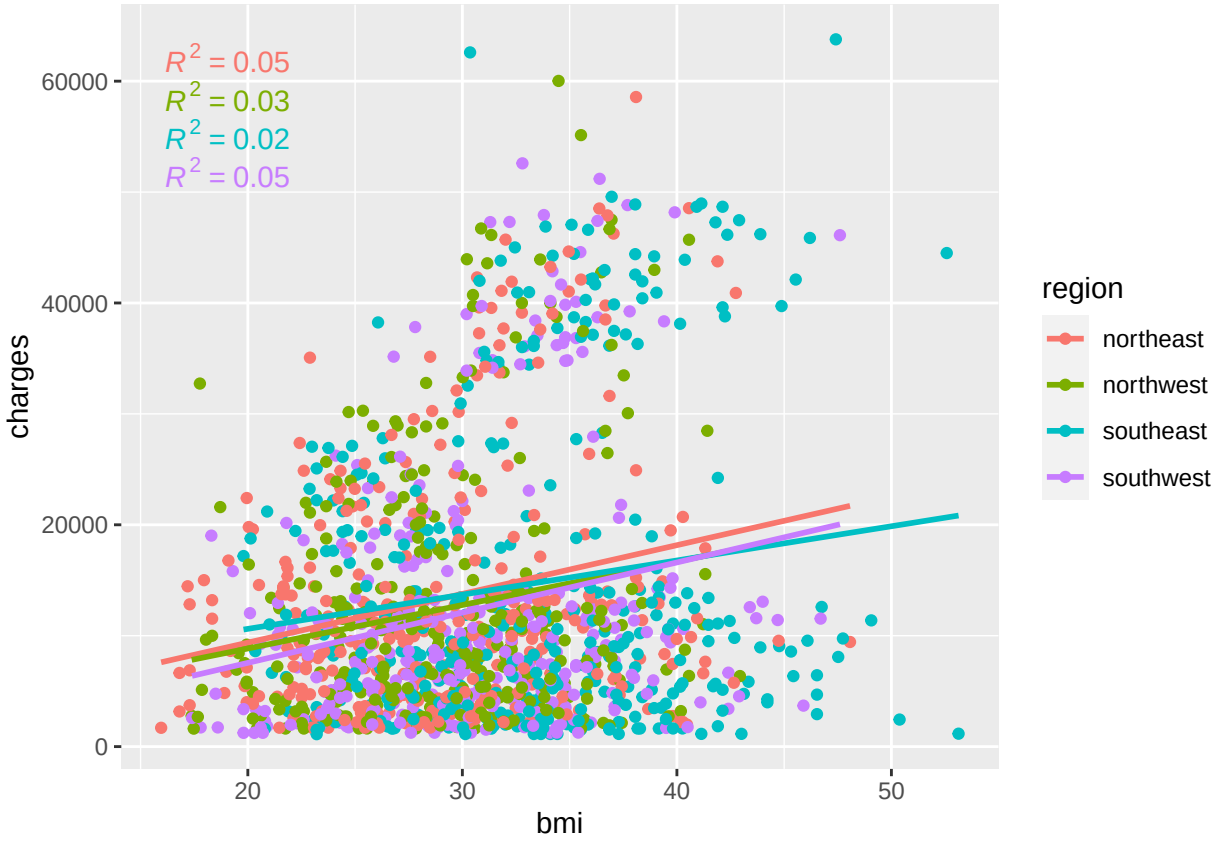

```
ggplot(data,aes(x=age,y=charges,fill=sex, color=sex))+geom_point()+ stat_poly_line() + stat_poly_eq()
```




```
ggplot(data,aes(x=age,y=charges,group=children, fill=children, color= children ))+geom_point()+ stat_pol
```

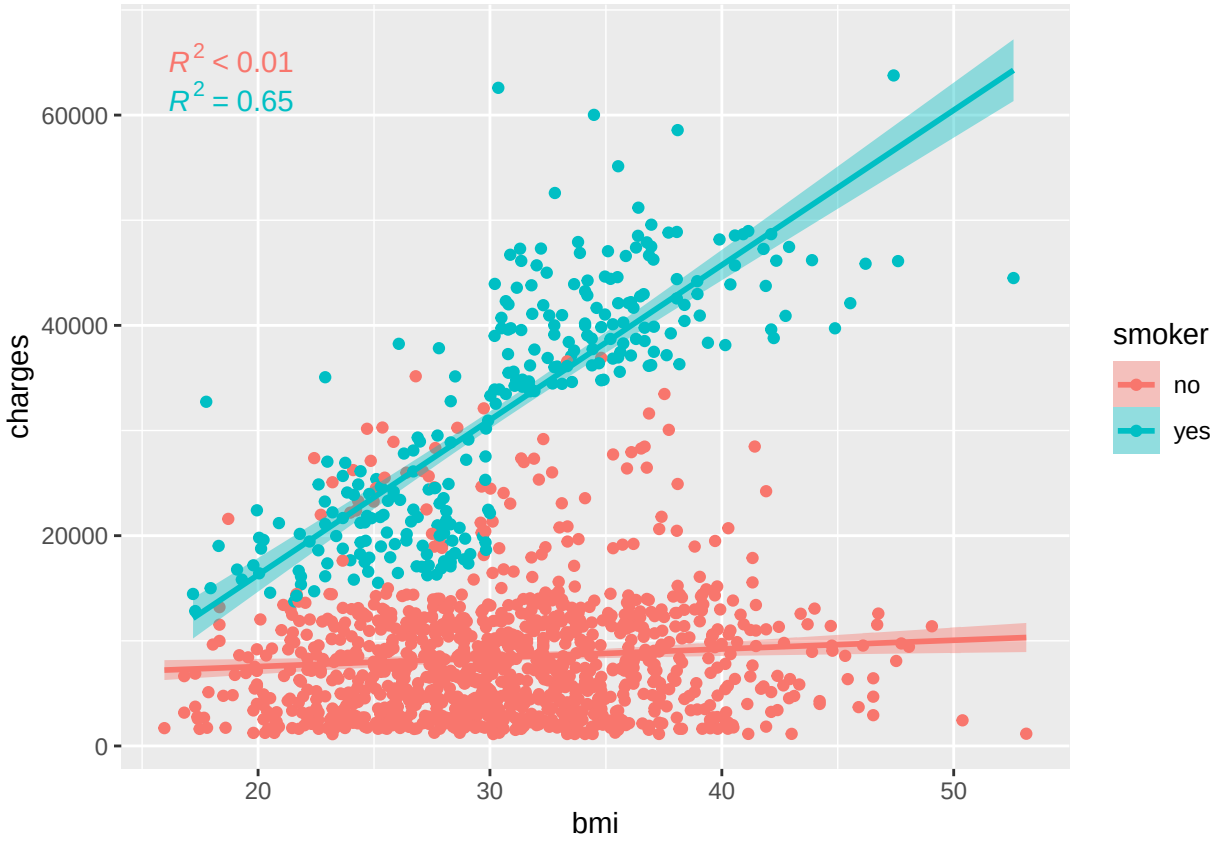



```
ggplot(data,aes(x=bmi,y=charges,fill=region, color=region))+geom_point()+stat_poly_line(se=FALSE) + sta
```

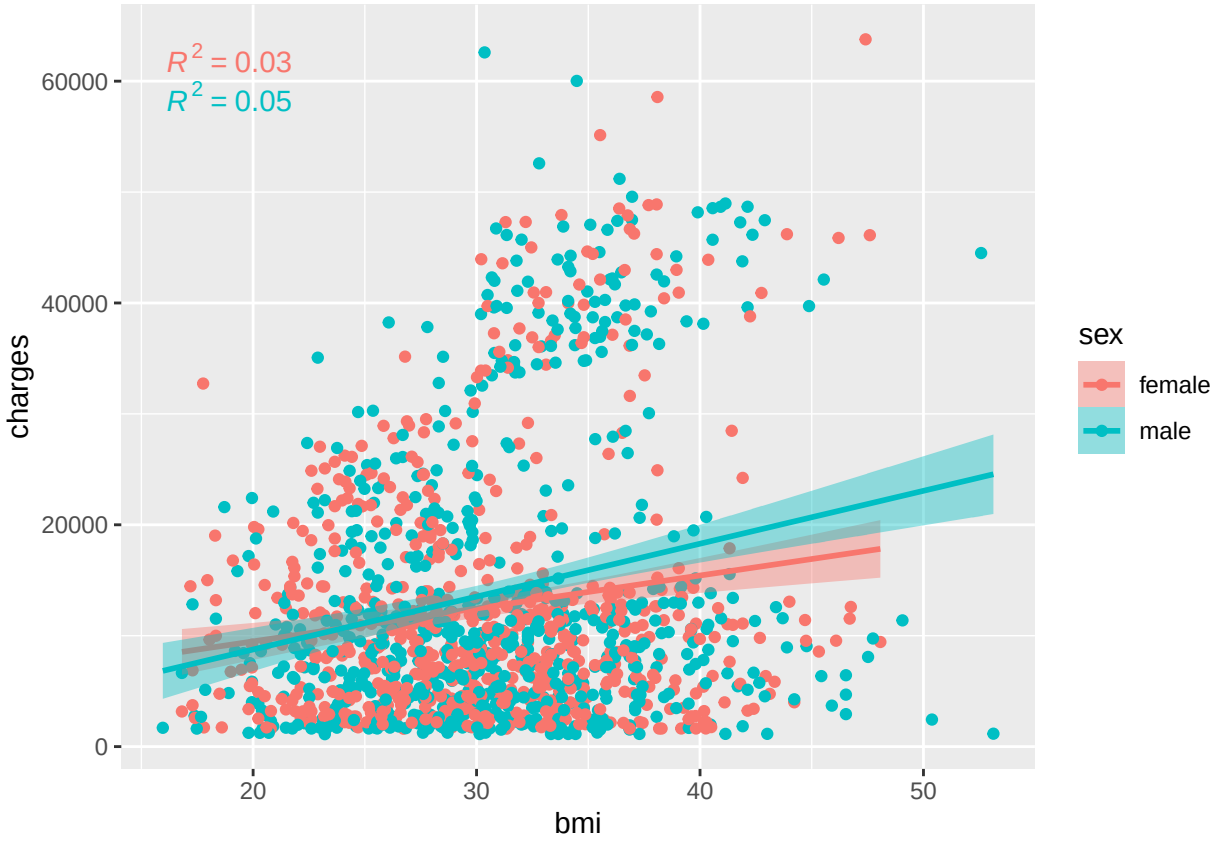


the scatter plots of charges vs. age, with each plot filled by a categorical variable such as "smoker","sex", "children" and "region" The group of individuals who is smoker or having 5 children appears to have a high R-squared value.

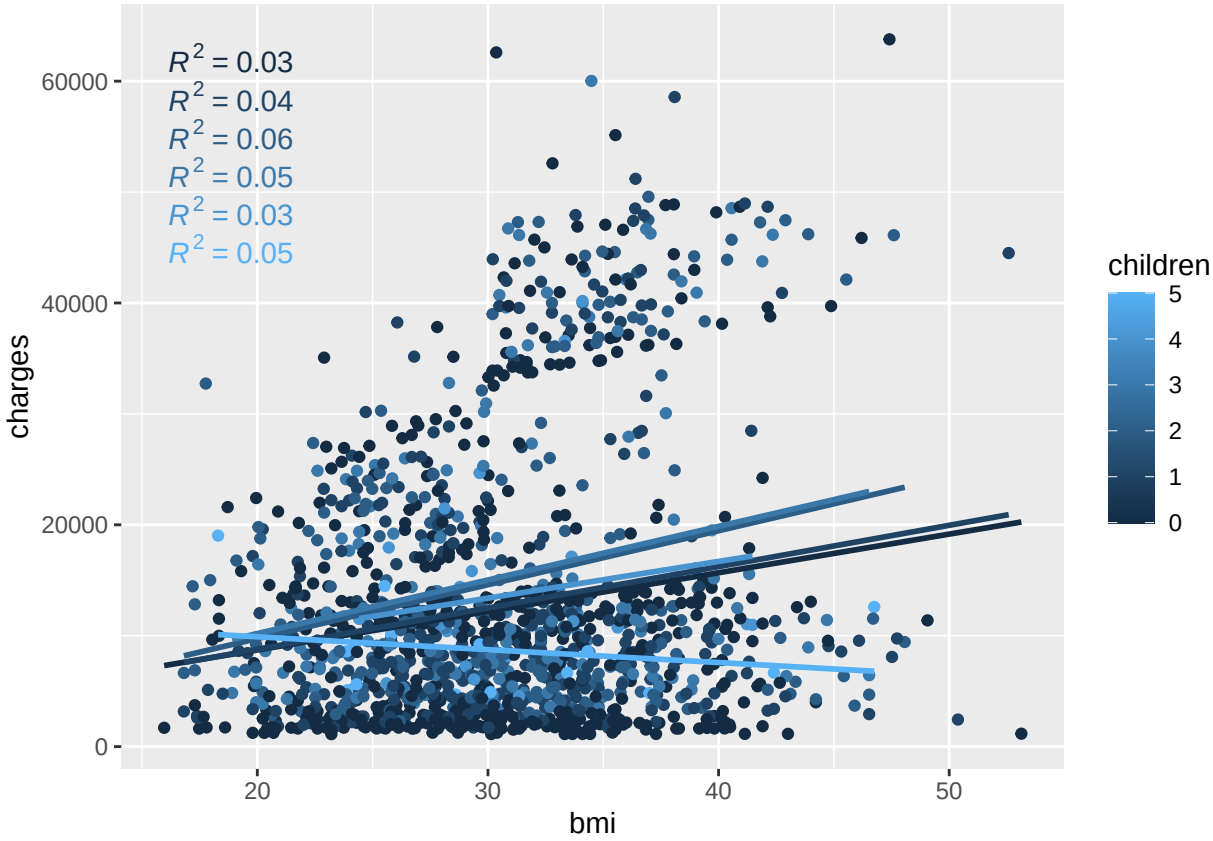
```
ggplot(data,aes(x=bmi,y=charges,fill=smoker, color=smoker))+geom_point()+ stat_poly_line() + stat_poly_
```



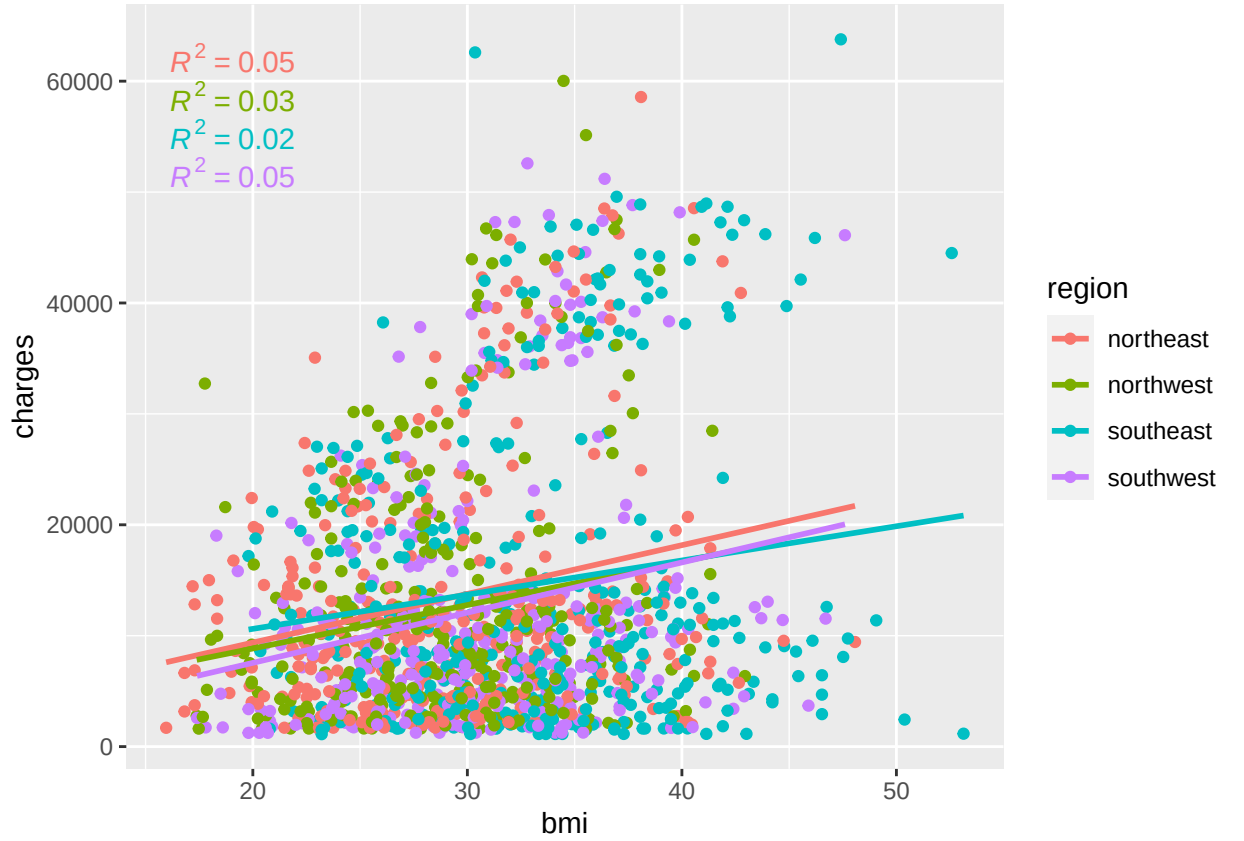

```
ggplot(data,aes(x=bmi,y=charges,fill=sex, color=sex))+geom_point()+ stat_poly_line() + stat_poly_eq()
```




```
ggplot(data,aes(x=bmi,y=charges,group=children, fill=children, color= children ))+geom_point()+ stat_pol
```




```
ggplot(data,aes(x=bmi,y=charges,fill=region, color=region))+geom_point()+stat_poly_line(se=FALSE) + sta
```



the scatter plots of charges vs. bmi, with each plot filled by a categorical variable such as "smoker","sex", "children" and "region" The group of individuals who is smoker appears to have a high R-squared value.

multiple linear regression

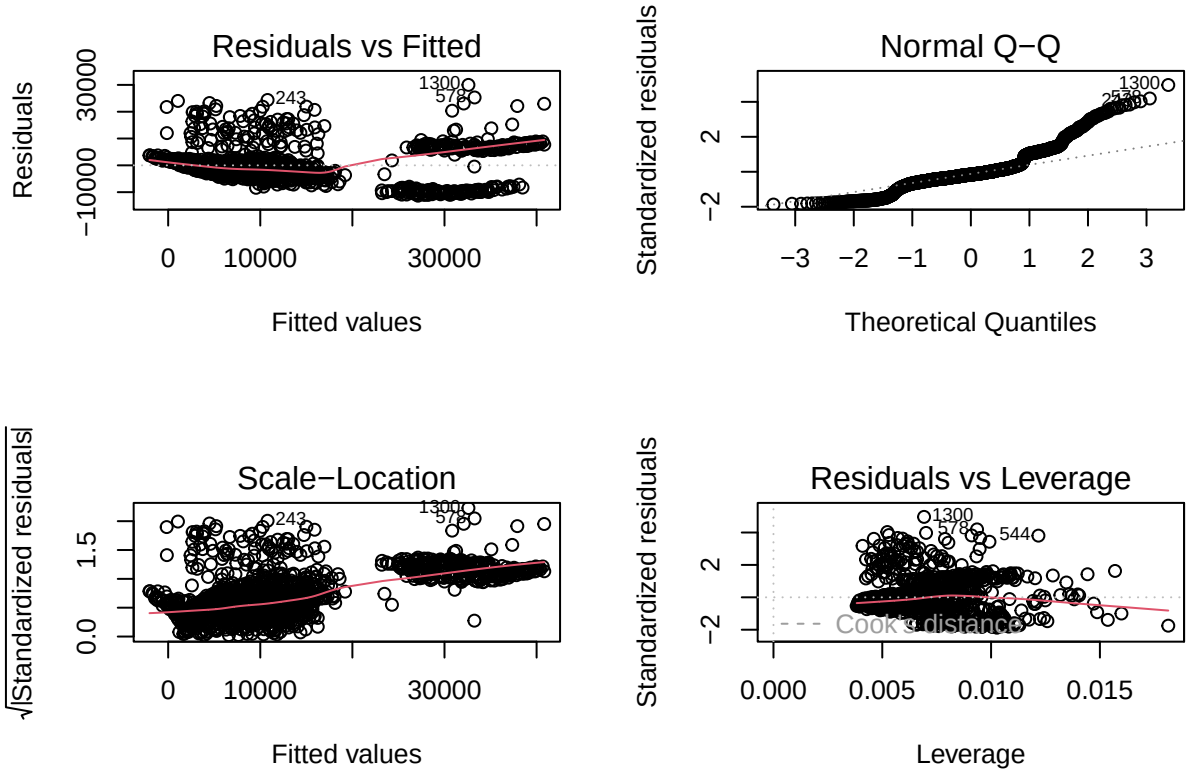
```
fit <- lm(charges ~ age + sex + bmi + children + smoker + region, data=data)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11305.1  -2850.3   -979.9   1395.0  29992.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11936.56     988.23 -12.079  < 2e-16 ***
## age                256.76      11.91  21.555  < 2e-16 ***
## sexmale           -129.48     333.20  -0.389 0.697630
## bmi                339.25      28.61  11.857  < 2e-16 ***
## children           474.82     137.90   3.443 0.000593 ***
## smokeryes        23847.33     413.35  57.693  < 2e-16 ***
```

```
## regionnorthwest   -349.23     476.82  -0.732 0.464053
## regionsoutheast  -1035.27     478.87  -2.162 0.030804 *
## regionsouthwest   -960.08     478.11  -2.008 0.044836 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6064 on 1328 degrees of freedom
## Multiple R-squared:  0.7507, Adjusted R-squared:  0.7492
## F-statistic:   500 on 8 and 1328 DF,  p-value: < 2.2e-16
```

age, bmi children, smokeryes, southeast and southwest are the variables or levels which apper to have significant impact on the insurance charges. However, we must check the assumption of this multiple linear regession before making conclusion.

```
par(mfrow=c(2,2))
plot(fit)
```




```
residual <- residuals(fit)
shapiro.test(residual) # check for normality assumption.
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residual
## W = 0.89909, p-value < 2.2e-16
```

```
vif(fit)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.016794  1        1.008362
## sex      1.008944  1        1.004462
## bmi      1.106742  1        1.052018
## children 1.004017  1        1.002006
## smoker   1.012100  1        1.006032
## region   1.099037  3        1.015864
```

Assumption:
Linearity: The residual vs. fitted plot indicates that the residuals are not randomly scattered around the horizontal line of zero. This indicates that the linear model may not be the best fit for the data. Not satisfy.

Normality: the Normal Q-Q plot and shapiro test indicate the residuals are not normally distributed. Not satisfy.

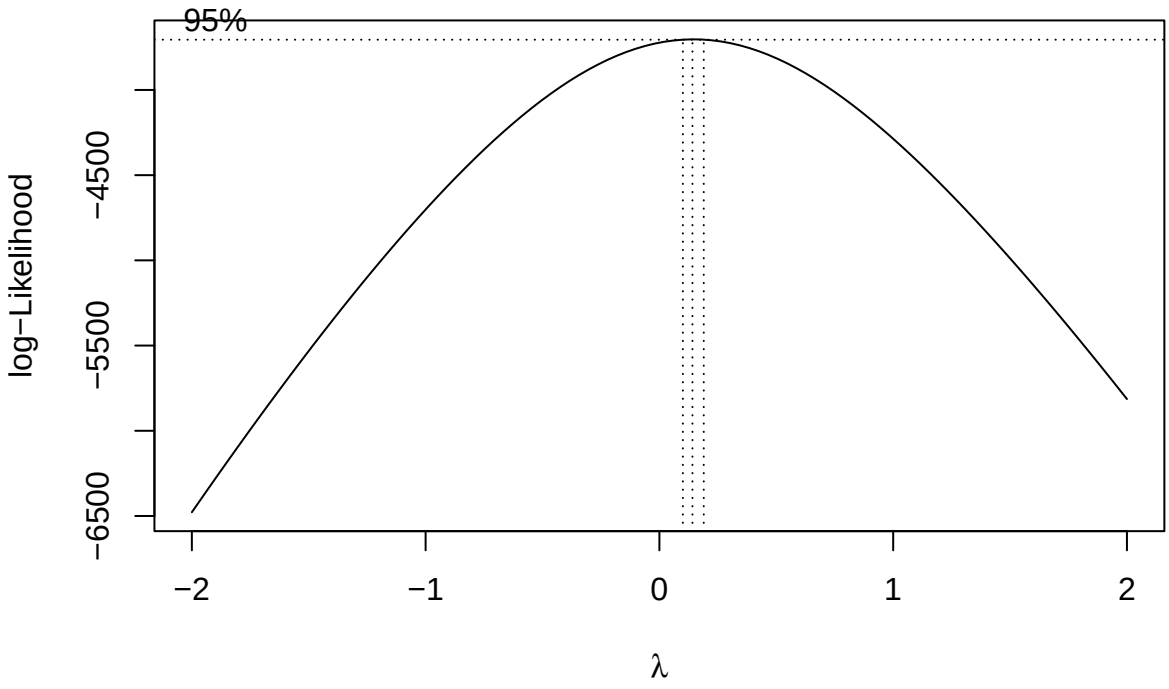
No multicollinearity: vif test shows there are no multicollinearity in the model, which is good.

Homogeneity: scale-location plot shows red line is not horizontal with points spread equally across the plot. This indicates heteroscedasticity exists.

Hence, assumptions of linear model are clearly not satisfied. I'll try to transform the dependent variable (charges) using box-cox transformation method to see whether the assumptions of linear model would be satisfied with transformed variable.

box-cox transformation

```
bc <- boxcox(charges ~ age + sex + bmi + children + smoker + region , data=data)
```




```
(lambda <- bc$x[which.max(bc$y)])
```

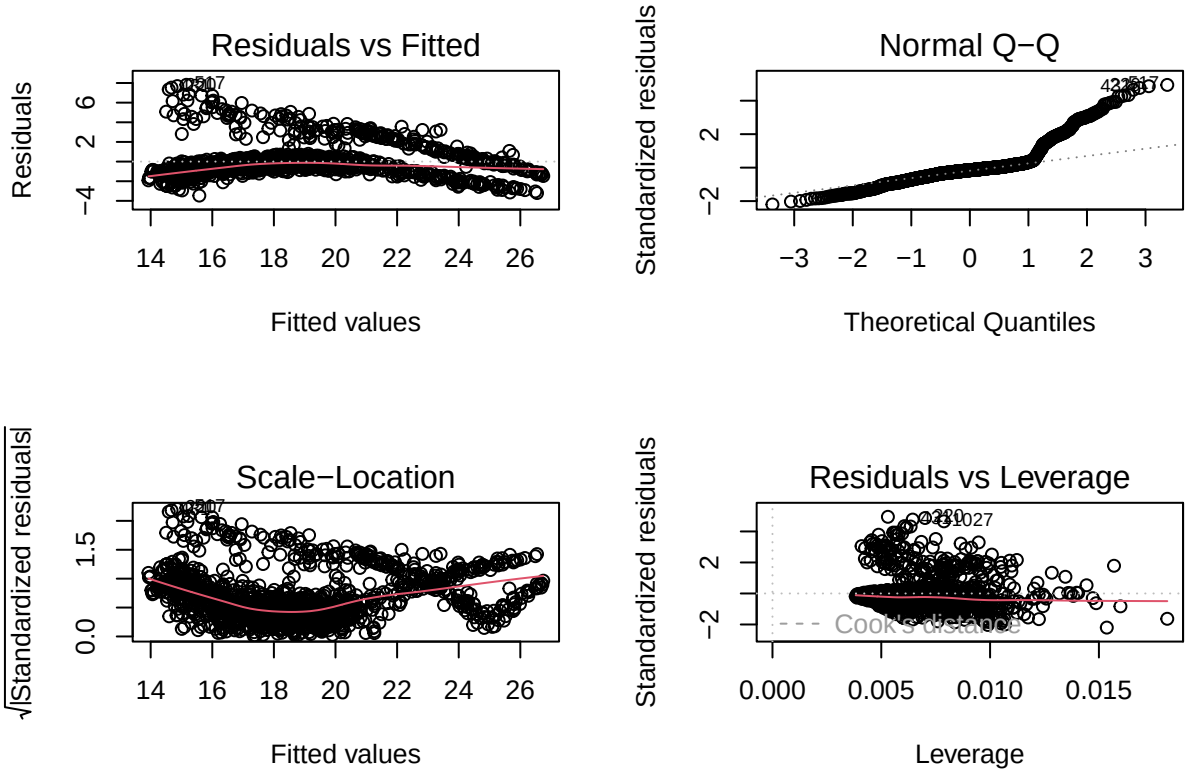
```
## [1] 0.1414141
```

```
new_model <- lm(((charges^lambda-1)/lambda) ~ age + sex + bmi + children + smoker + region, data=data)
summary(new_model)
```

```
##
## Call:
## lm(formula = ((charges^lambda - 1)/lambda) ~ age + sex + bmi +
##     children + smoker + region, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4625 -0.7628 -0.2405  0.1752  7.8156
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.272017   0.258363  43.629  < 2e-16 ***
## age              0.118258   0.003114  37.972  < 2e-16 ***
## sexmale         -0.228318   0.087111  -2.621 0.008867 **
## bmi              0.055196   0.007480   7.379 2.80e-13 ***
## children         0.326474   0.036052   9.056  < 2e-16 ***
## smokeryes        5.895636   0.108066  54.556  < 2e-16 ***
```

```
## regionnorthwest -0.209865   0.124661  -1.683 0.092517 .
## regionsoutheast -0.525042   0.125195  -4.194 2.93e-05 ***
## regionsouthwest -0.439757   0.124996  -3.518 0.000449 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.585 on 1328 degrees of freedom
## Multiple R-squared:  0.7759, Adjusted R-squared:  0.7745
## F-statistic: 574.7 on 8 and 1328 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(new_model)
```



Transforming variable does not to satisfy the assumptions of linear model. Hence, we should perform decision tree-based method like random forest.

Random forest regression

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## 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:ggplot2':
##
##     margin
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(123)
training.samples <- data$charges %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- data[training.samples, ]
test.data <- data[-training.samples, ]
```

split dataset into train and test data, with 80% being train data and 20% being test data.

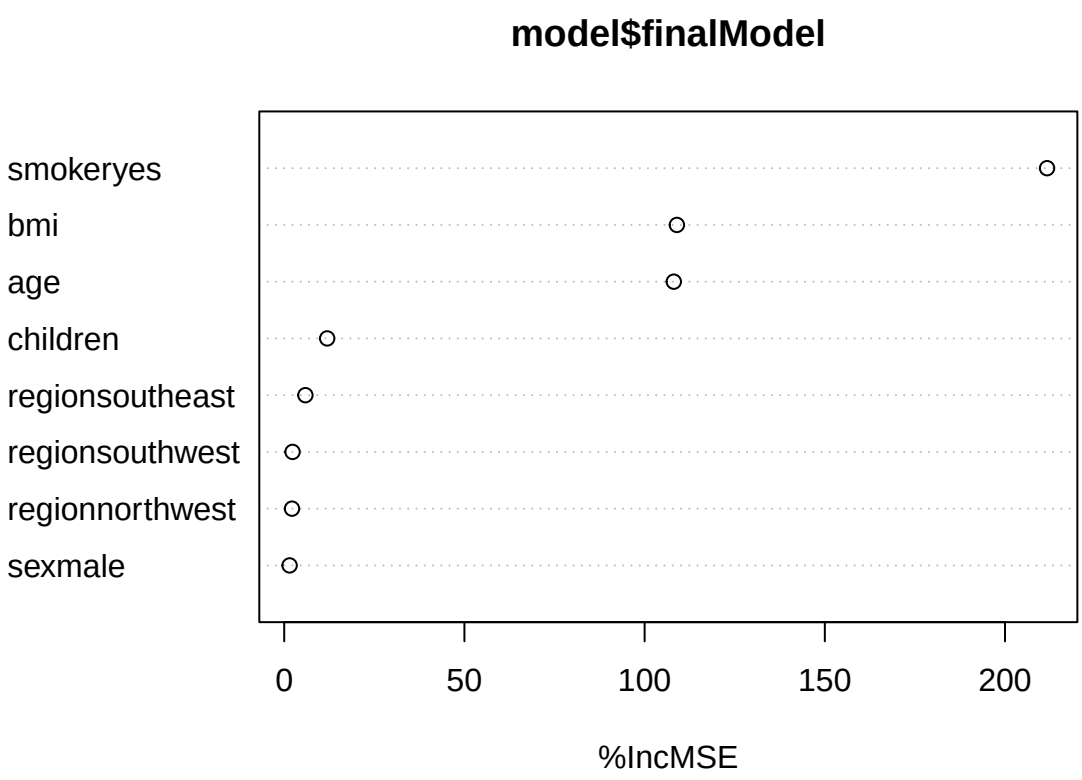
```
set.seed(123)
model <- train(
  charges ~., data = train.data, method = "rf",
  trControl = trainControl("cv", number = 10),
  importance = TRUE
  )
```

10-fold cross-validation would be used

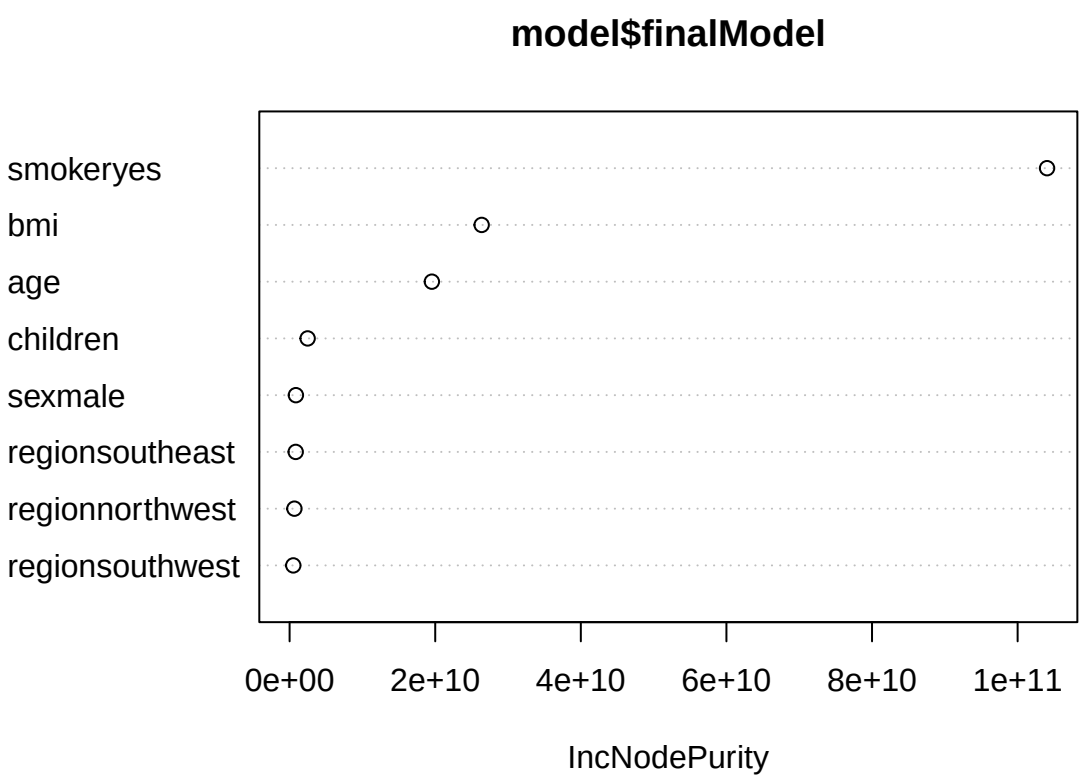
```
model$bestTune #best set of tuning parameter.
```

```
##   mtry
## 2    5
```

```
varImpPlot(model$finalModel, type = 1)
```




```
varImpPlot(model$finalModel, type = 2)
```




```
varImp(model)
```

```
## rf variable importance
##
##                   Overall
## smokeryes       100.0000
## bmi              51.1292
## age              50.7188
## children          4.9639
## regionsoutheast   2.0844
## regionsouthwest   0.3861
## regionnorthwest   0.3150
## sexmale           0.0000
```

Smokeryes is the most significant variable in determining insurance charges, followed by BMI and age. Other factors have minor or no impact on the charges

```
predictions <- model %>% predict(test.data)
head(predictions)
```

```
##         5         9        14        19        25        28
##  4877.633  7903.622 12120.191 13150.115  6576.011 12628.450
```

```
RMSE(predictions, test.data$charges)
```

```
## [1] 4806.051
```

The root mean squared error between the predicted charges and the actual charges in the test data is 4806.051.