QBS121_FINAL_PROJECT

GROUP3

2/26/2022

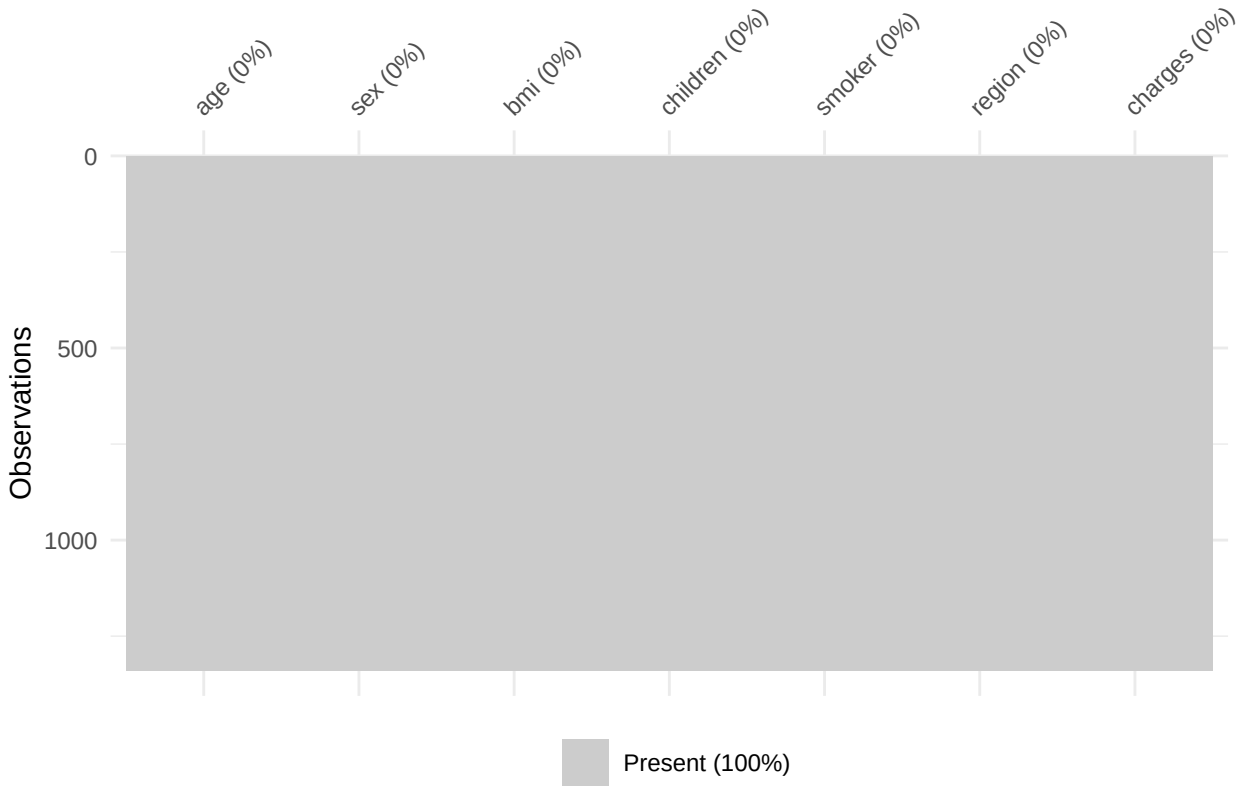
```
#reading the dataset
insurance<-read.csv("/Users/yoo/Downloads/insurance.csv")
```

```
#looking into the variables
str(insurance)
```

```
## '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 ...
```

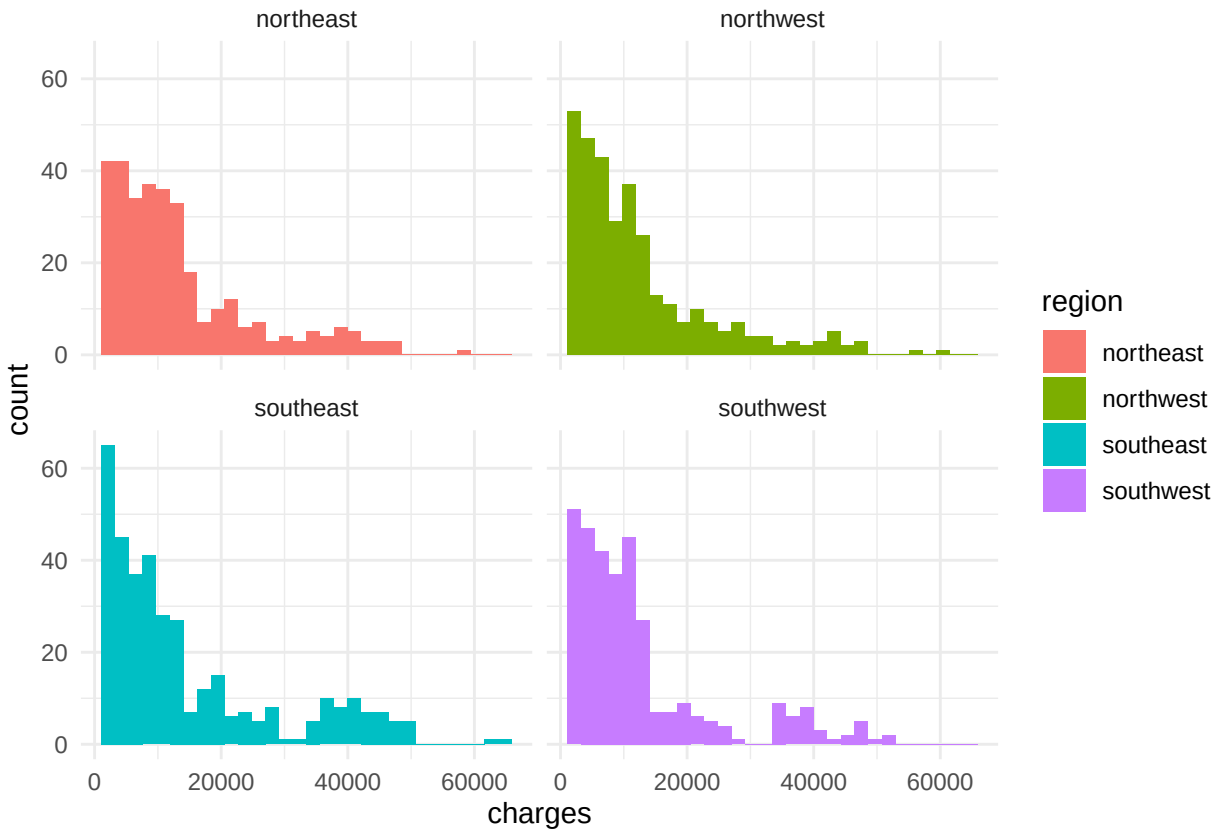
```
#converting the str into factor types
insurance$smoker<-as.factor(insurance$smoker)
insurance$sex<-as.factor(insurance$sex)
insurance$region<-as.factor(insurance$region)
```

```
#checking for missing mess in the dataset.
library(visdat)

insurance %>%
  visdat::vis_miss()
```



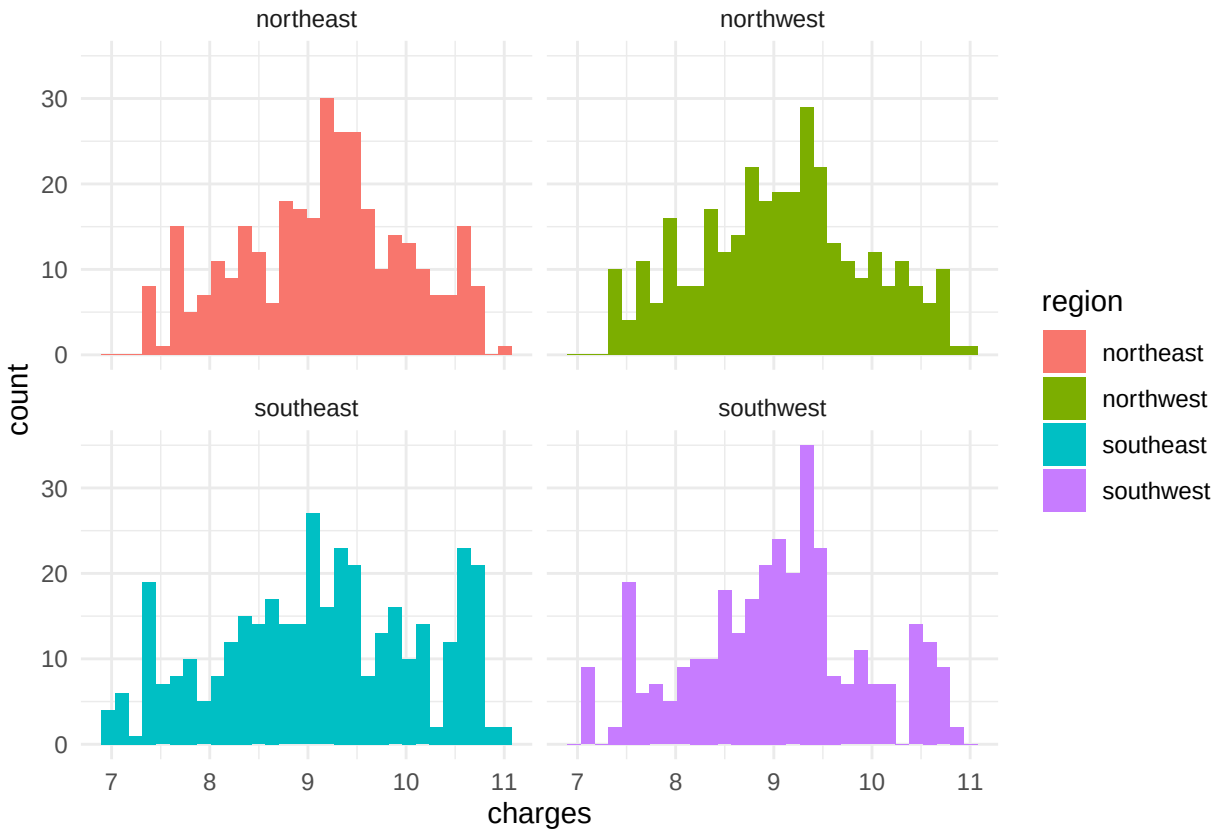

```
#we are plotting the dependent variable
#from this we can concluse that it is not normally distributed
insurance %>%
  as_tibble() %>%
  select(region, charges) %>%
  ggplot(aes(charges, fill = region)) +
  geom_histogram() +
  facet_wrap(~region) +
  theme(legend.position = "none") +
  theme_minimal()
```

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




```
insurance %>%
  as_tibble() %>%
  mutate(charges = log(charges)) %>%
  select(region, charges) %>%
  ggplot(aes(charges, fill = region)) +
  geom_histogram() +
  facet_wrap(~region) +
  theme(legend.position = "none") +
  theme_minimal()#to make the linear regression result reliable, the target var should be normally dist
```

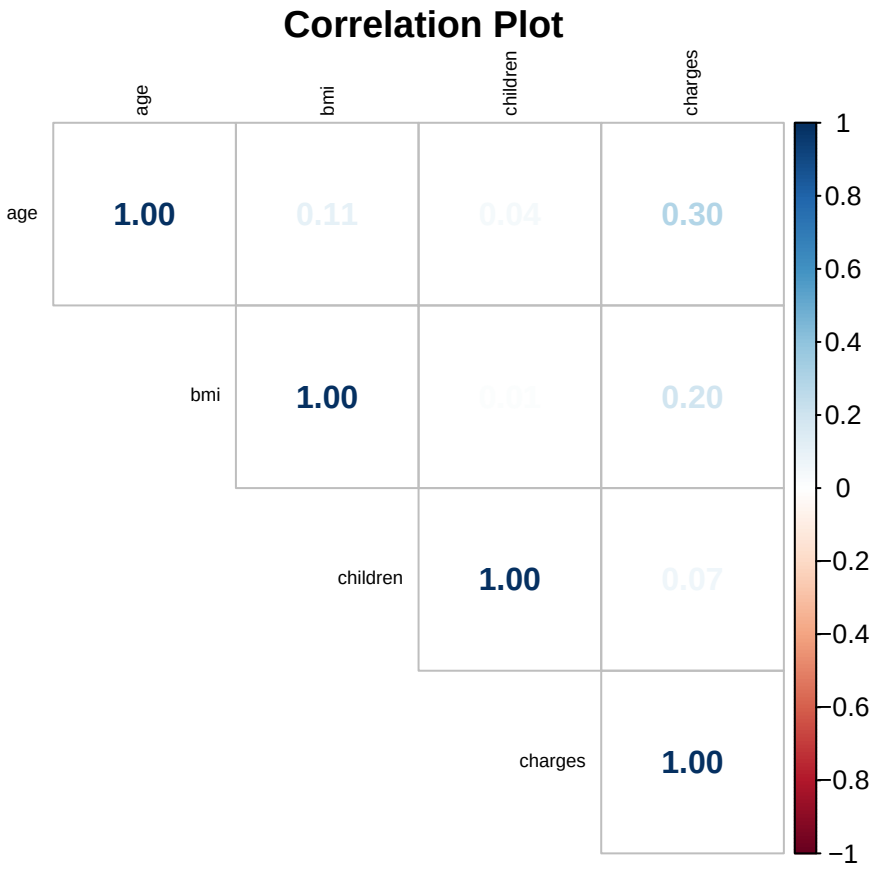
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#checking corelation between the variables
cor(insurance$charges,insurance$age)#0.2990082
```

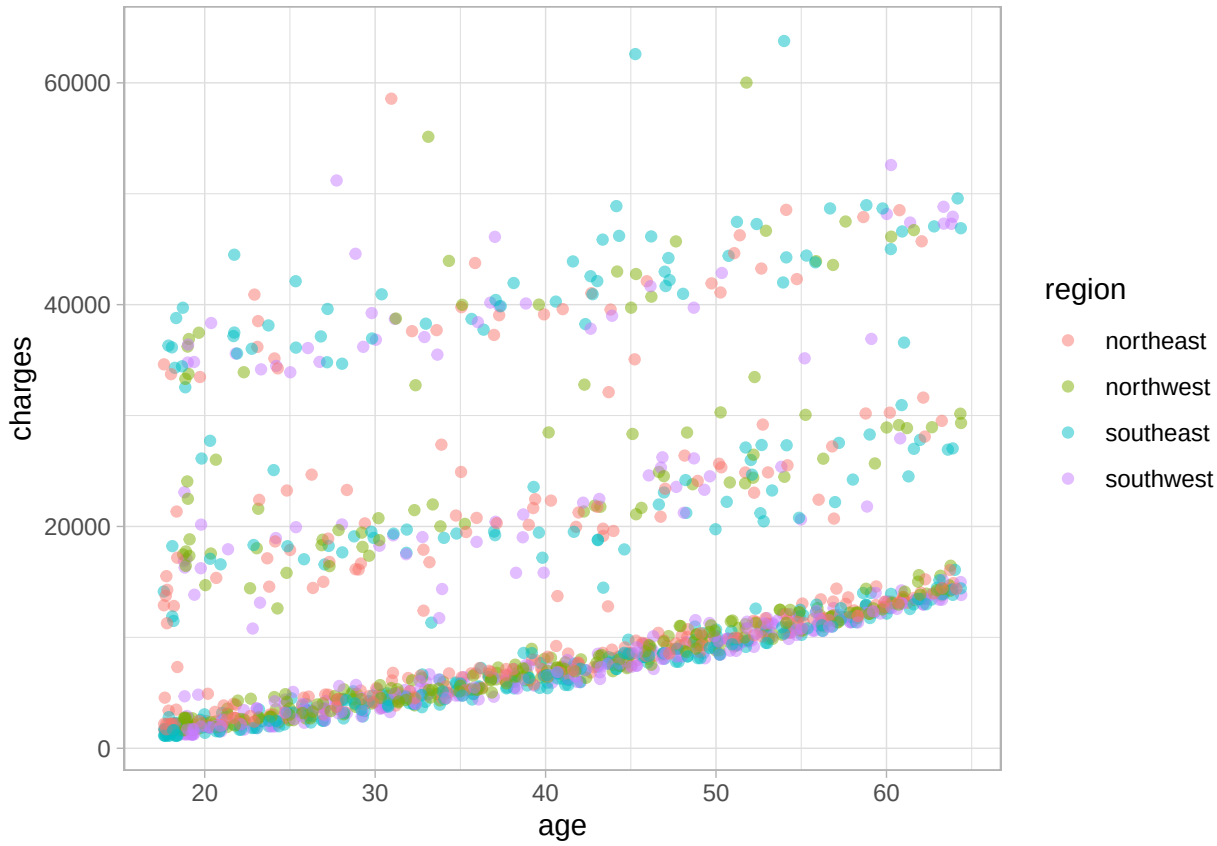
```
## [1] 0.2990082
```

```
insurance_n <- select_if(insurance, is.numeric)
corrmatrix <- cor(insurance_n)
corrplot::corrplot(corrmatrix, method=c("number"), type = "upper",tl.cex=.6
, tl.col="black", title="Correlation Plot",number.font = 2, mar=c(0,0,1,0), )
```

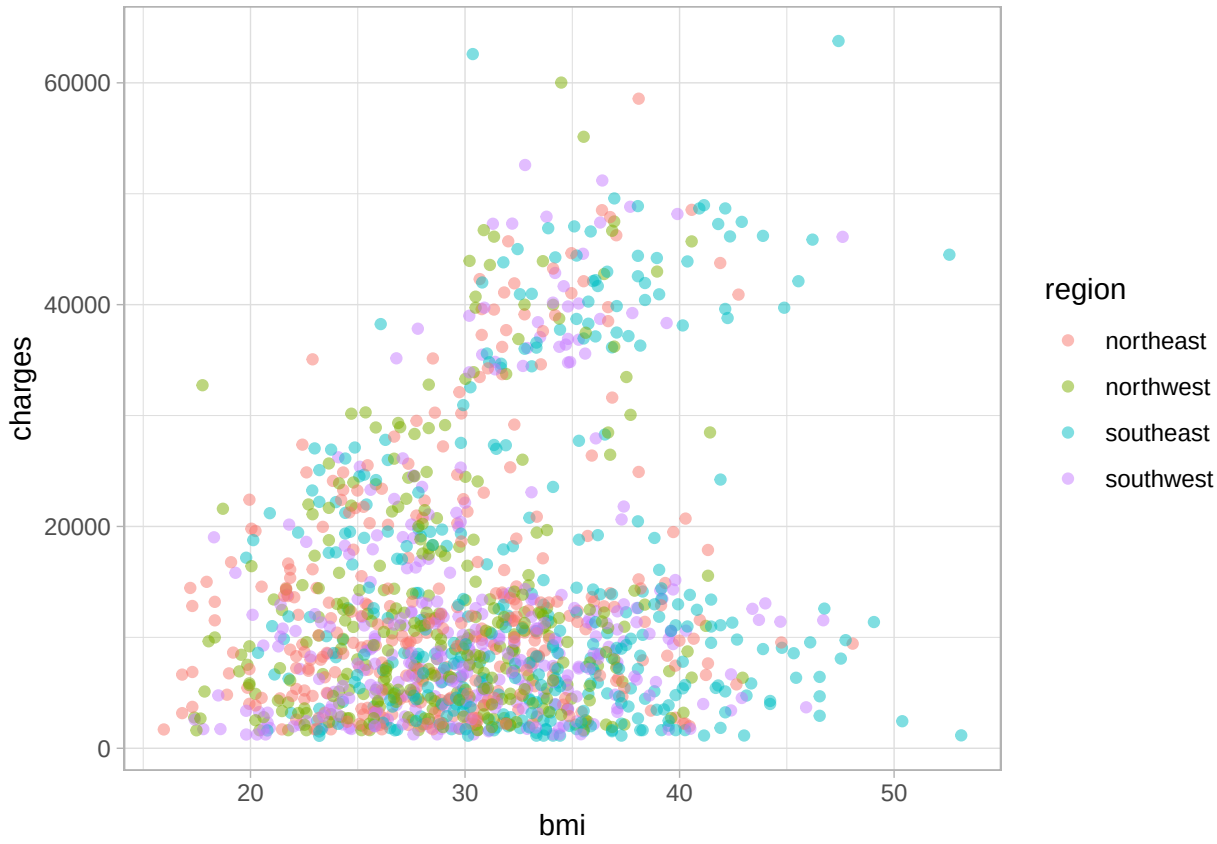



```
#from this we can say that charges comparitively highly corelated with "age" followed by "bmi". And the
```

```
#looking into the correlation between charges and age, charges and bmi based on the region

Data<-insurance
x <- ggplot(Data, aes(age, charges,color=region)) +
  geom_jitter( alpha = 0.5) +
    theme_light()

y <- ggplot(Data, aes(bmi, charges, color = region)) +
  geom_jitter( alpha = 0.5) +
  theme_light()

x
```

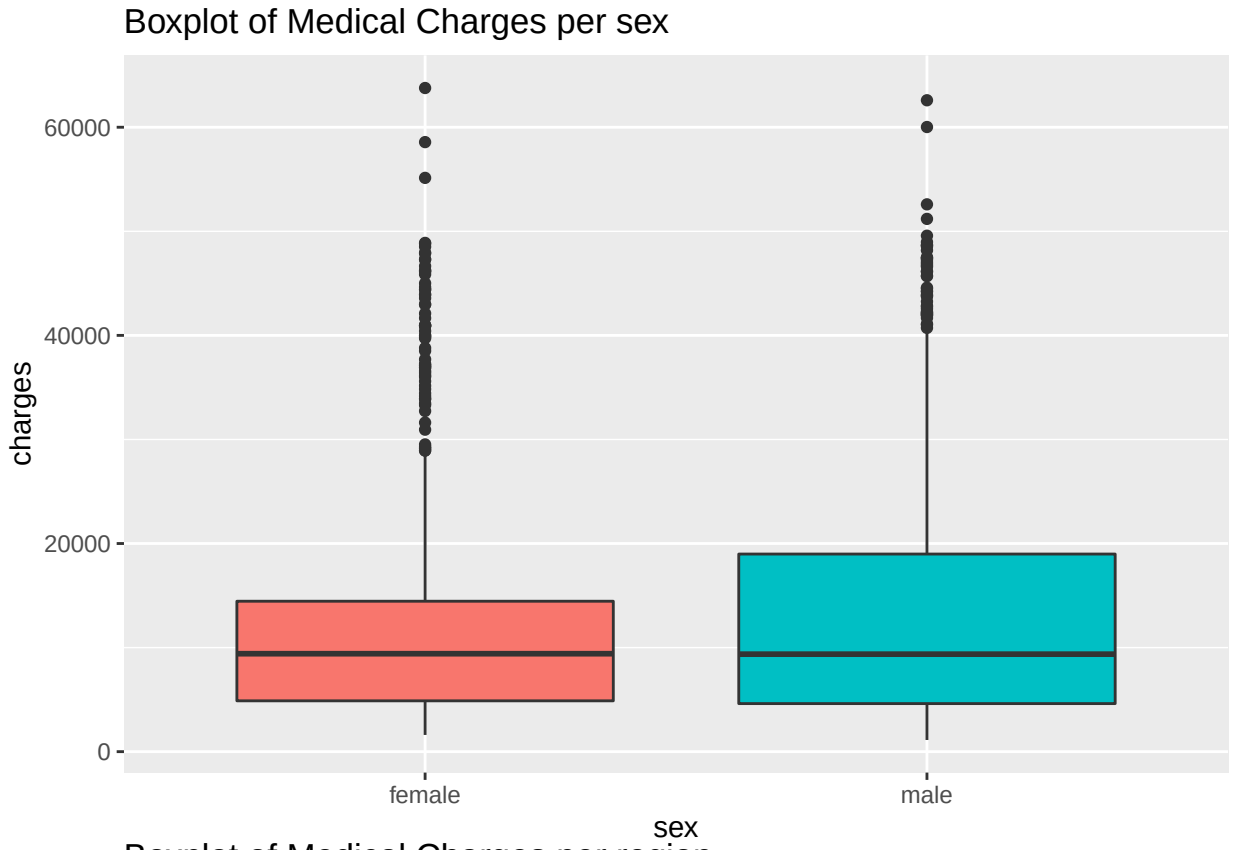

60000
40000
20000
0
charges
20
30
40
50
60
age
region
northeast
northwest
southeast
southwest

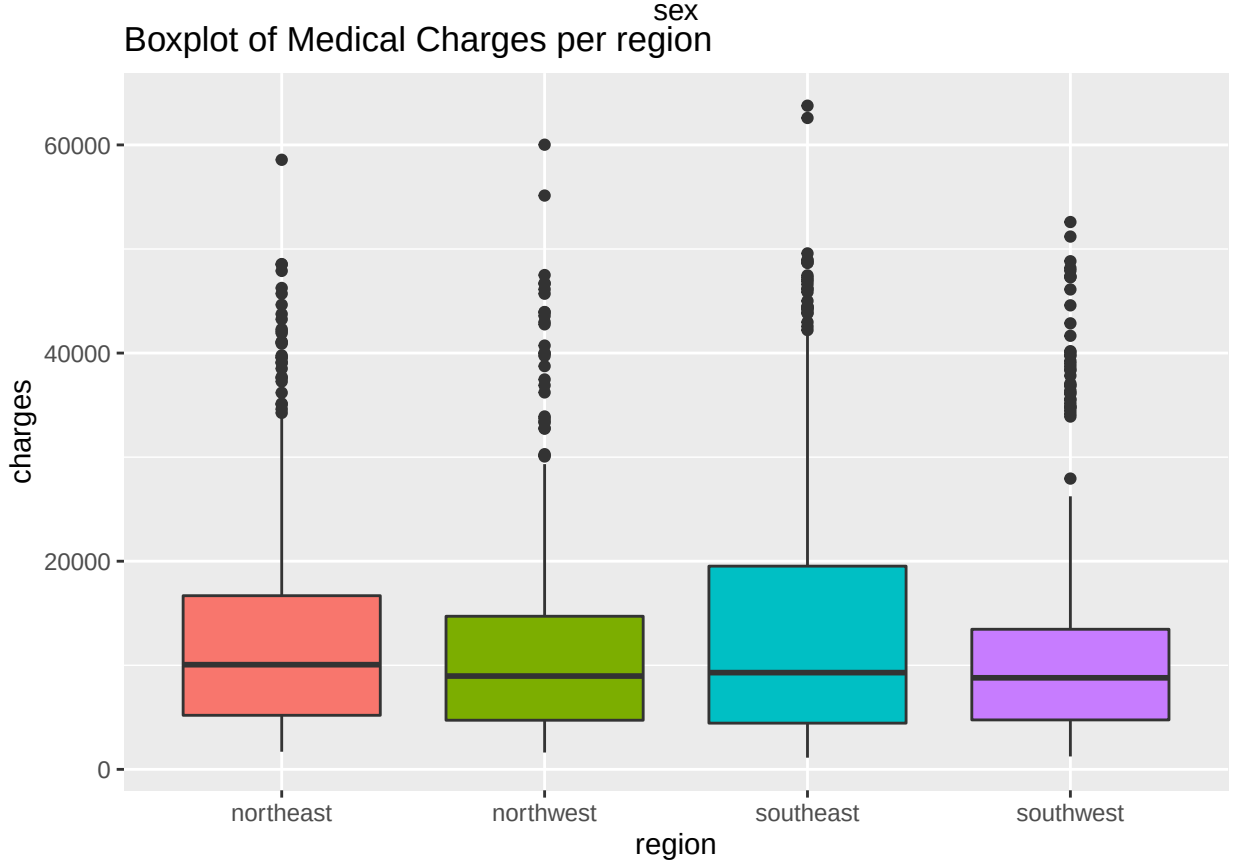
```
y
```

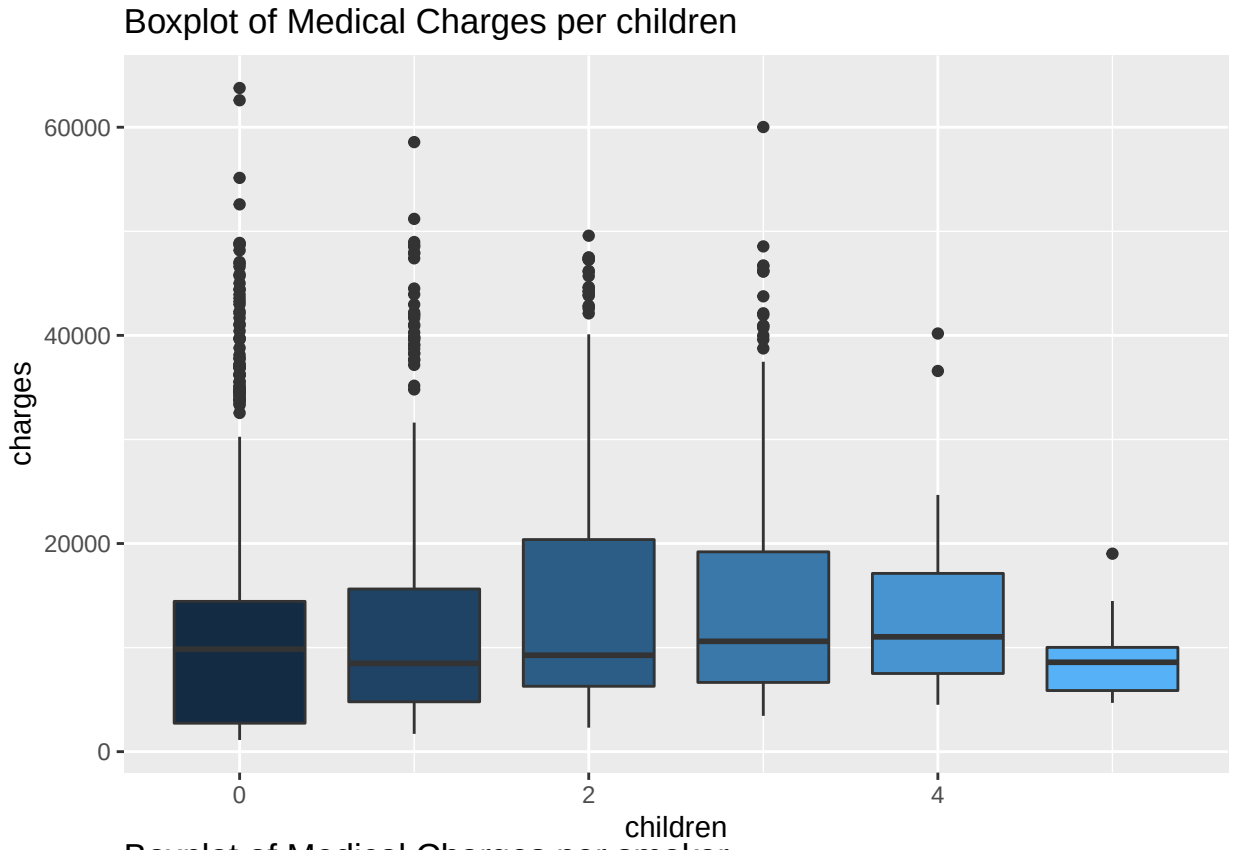


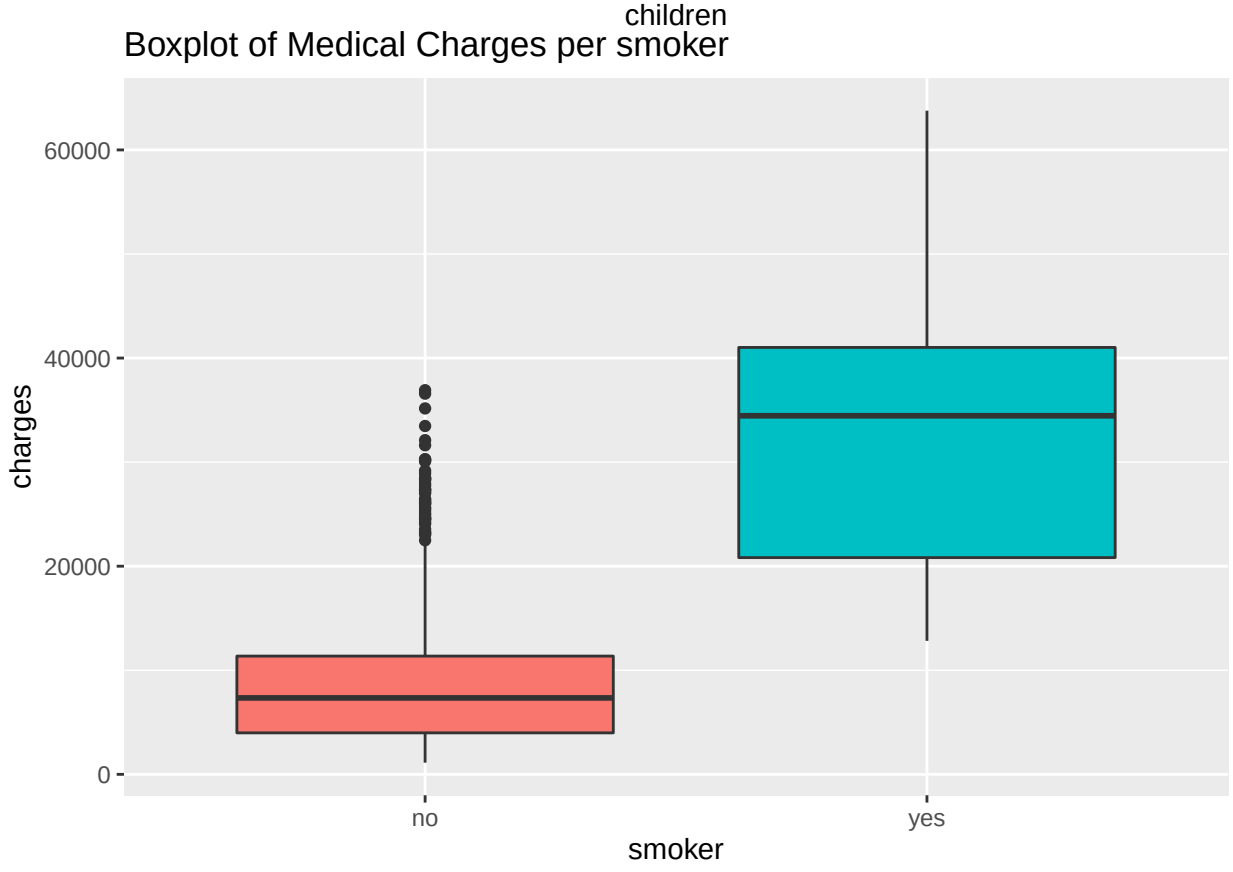

```
# p <- plot_grid(x, y)
# title <- ggdraw() + draw_label("1. Correlation between Charges and Age / BMI", fontface='bold')
# plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))



#As Age go up Charges for health insurance also trends up.
```

```
#plot between the dependent variable and the all the other independent variable
for (col in c('sex', 'region', 'children', 'smoker')) {
  plot <- ggplot(data = insurance,
                 aes_string(x = col, y = 'charges', group = col, fill = col)) +
            geom_boxplot(show.legend = FALSE) +
            ggtitle(glue::glue("Boxplot of Medical Charges per {col}"))
  print(plot)
}
```


Boxplot of Medical Charges per sex
charges
60000
40000
20000
0
female
male
sex


Boxplot of Medical Charges per region
charges
60000
40000
20000
0
northeast
northwest
southeast
southwest
region


Boxplot of Medical Charges per children
charges
60000
40000
20000
0
0
2
4
children


Boxplot of Medical Charges per smoker
charges
60000
40000
20000
0
no
yes
smoker

```
model1_linear<-lm(charges~age+bmi+region+children+sex+smoker,data=insurance)
summary(model1_linear)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + region + children + sex +
##     smoker, data = insurance)
##
## 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)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## bmi                339.2       28.6  11.860  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## children           475.5      137.8   3.451 0.000577 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## ---
## 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
```

```
#install.packages('sjPlot')
library(sjPlot)
```

```
## #refugeeswelcome
```

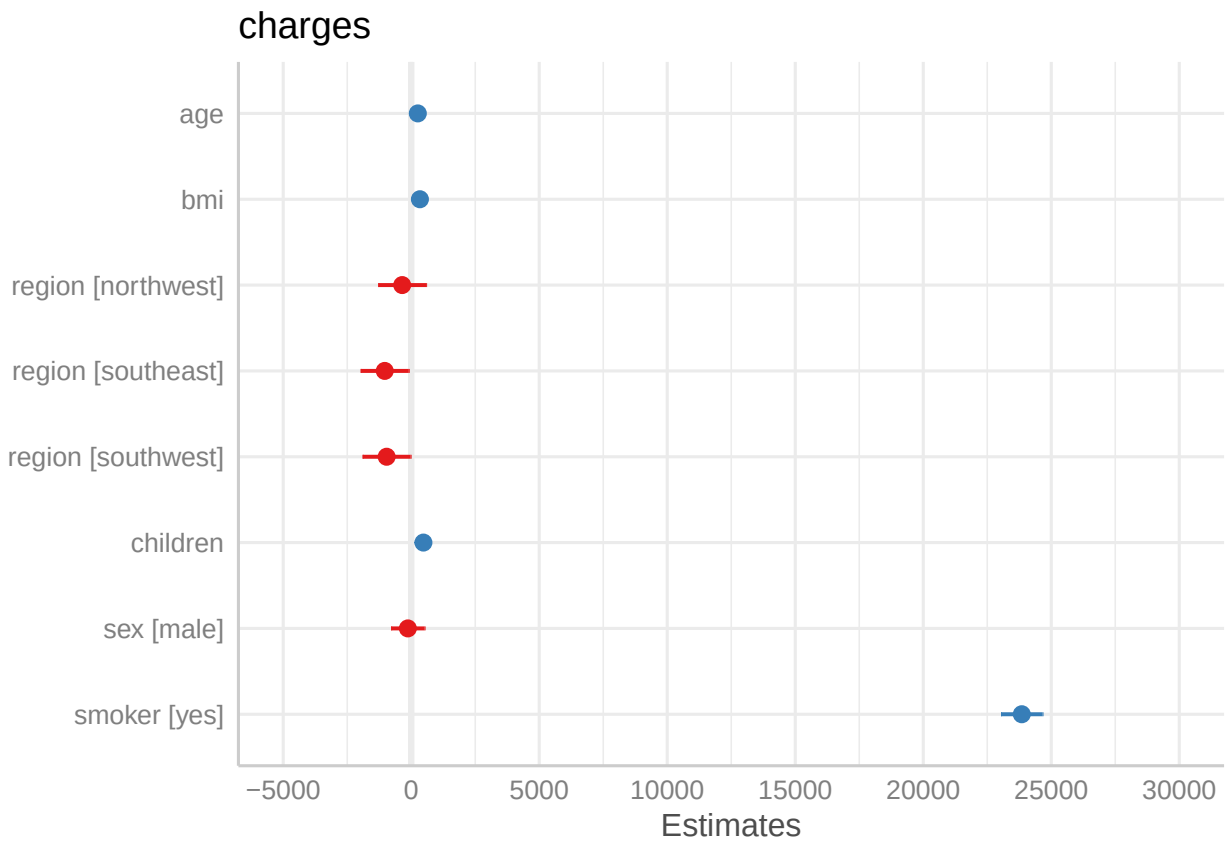
```
library(sjlabelled)
```

```
##
## Attaching package: ’sjlabelled’

## The following object is masked from ’package:forcats’:
##
##     as_factor

## The following object is masked from ’package:dplyr’:
##
##     as_label

## The following object is masked from ’package:ggplot2’:
##
##     as_label
```

```
library(sjmisc)
```

```
##
## Attaching package: 'sjmisc'

## The following object is masked from 'package:purrr':
##
##     is_empty

## The following object is masked from 'package:tidyr':
##
##     replace_na

## The following object is masked from 'package:tibble':
##
##     add_case
```

```
library(ggplot2)

theme_set(theme_sjplot())
plot_model(model1_linear)
```




```
model2_linear<-lm(charges~age+bmi+smoker,data=insurance)
summary(model2_linear)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker, data = insurance)
##
## 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 ***
## age            259.55      11.93   21.75   <2e-16 ***
## bmi            322.62      27.49   11.74   <2e-16 ***
## smokeryes    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
```

```
insurance$charges_cat <- ifelse(insurance$charges >median(insurance$charges), 1, 0)
```

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.1.2
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
model_glm<-glm(charges_cat~age+sex+bmi+children+smoker+region,family=binomial,data=insurance)
summary(model_glm)
```

```
##
## Call:
## glm(formula = charges_cat ~ age + sex + bmi + children + smoker +
##     region, family = binomial, data = insurance)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5354  -0.4102  -0.0423   0.4005   3.2846
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -8.17993    0.67948 -12.038  < 2e-16 ***
```

```
## age               0.16683    0.01004  16.624  < 2e-16 ***
## sexmale          -0.35313    0.18188  -1.942  0.05219 .
## bmi               0.03268    0.01582   2.065  0.03891 *
## children          0.14483    0.07495   1.932  0.05333 .
## smokeryes        22.32977  509.88463   0.044  0.96507
## regionnorthwest  -0.41109    0.25915  -1.586  0.11267
## regionsoutheast  -0.86119    0.26801  -3.213  0.00131 **
## regionsouthwest  -0.77646    0.25872  -3.001  0.00269 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1854.86  on 1337  degrees of freedom
## Residual deviance:  773.45  on 1329  degrees of freedom
## AIC: 791.45
##
## Number of Fisher Scoring iterations: 18
```

```
model_lmer<-lmer(charges~sex+age+children+smoker+(1|region),data=insurance)
summary(model_lmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: charges ~ sex + age + children + smoker + (1 | region)
##    Data: insurance
##
## REML criterion at convergence: 27175.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4271 -0.3074 -0.2043 -0.0667  4.5797
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  region   (Intercept)    37525  193.7
##  Residual             40605989 6372.3
## Number of obs: 1338, groups:  region, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) -2887.16     579.96  -4.978
## sexmale        59.49     349.57   0.170
## age           273.18      12.42  21.992
## children      488.65     144.75   3.376
## smokeryes   23822.02     433.37  54.970
##
## Correlation of Fixed Effects:
##           (Intr) sexmal age    chldrn
## sexmale   -0.305
## age       -0.838  0.020
## children  -0.231 -0.018 -0.043
## smokeryes -0.148 -0.075  0.024 -0.008
```

```
model_glmer<-glmer(charges_cat~sex+age+children+smoker+(1|region),family=binomial,data=insurance)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00217927 (tol = 0.002, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
summary(model_glmer)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: charges_cat ~ sex + age + children + smoker + (1 | region)
##    Data: insurance
##
##      AIC      BIC   logLik deviance df.resid
##    797.8    829.0   -392.9    785.8     1332
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4753 -0.2996 -0.0371  0.2940 12.3929
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  region (Intercept) 0.05716  0.2391
## Number of obs: 1338, groups:  region, 4
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.707e+00  5.044e-01 -15.279   <2e-16 ***
## sexmale     -3.338e-01  1.805e-01  -1.849   0.0644 .
## age          1.671e-01  9.990e-03  16.727   <2e-16 ***
## children     1.440e-01  7.437e-02   1.937   0.0528 .
## smokeryes    2.255e+05  3.620e+02 622.982   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) sexmal age    chldrn
## sexmale   -0.102
## age       -0.921 -0.076
## children  -0.365 -0.012  0.198
## smokeryes  0.000  0.000  0.000  0.000
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00217927 (tol = 0.002, component 1)
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```